An Exploration of the Change of Teacher Assessment Practice, in Physical Education at Key Stage 3, between 2000 and 2005/2006

Thesis

Diane Burkinshaw

An exploration of the change of teacher assessment practice, in Physical Education at Key Stage 3, between 2000 and 2005 / 2006.

DOCTOR OF EDUCATION (EdD)

April 2011

DATE OF SUBMISSION: 31 OCT 2009
DATE OF AWARD : 17 MAY 2011

Abstract

Teacher assessment has been the 'modus operandi' in Physical Education (PE) since its inclusion as a foundation subject in the National Curriculum in 1992, yet the Office for Standards in Education (Ofsted) has consistently reported that assessment in this subject is problematic (1995, 1998, 2003, 2009). This research focuses on schools involved in initial teacher education and training, in partnership with Riverside University. Using an overarching case study strategy and a mixed methods approach to data collection and analysis (Yin, 2003), this longitudinal study explores the changes in teachers' assessment practice in PE at Key Stage 3, over a seven-year period, at a time of unprecedented reform of teacher assessment and its relationship with learning at national level (Black and Wiliam, 1998a; DfES 2004; ARG, 1999 –2010).

Using the work of Harlen (2004a) as a tool for analysis, it demonstrates that within the framework of NCPE (2000), at the three data collection points (2000, 2005 and 2006) PE teachers, in the Riverside Partnership, are using an ever-wider range of methods and tools, in order to make dependable assessment judgements at Key Stage 3. There is evidence that teacher assessment practice in PE has developed in line with current thinking at national level, particularly in terms of involving the pupils in their own assessment to inform their learning. However, teacher observation remains the dominant assessment mode.

The study concludes that driven by the prevailing culture of performativity and accountability (Broadfoot, 2000b; Ball, 2003) in the schools in which the teachers were working, the PE teachers' assessment practice increasingly moved towards the notion of 'good practice' in assessment in PE, as defined by Ofsted (2003b). However, given that teacher assessment practice continues to vary across the schools in the partnership, issues of consistency remain for the initial teacher training of the PE student teachers.

Acknowledgements

My thanks go firstly to my supervisor Dr. Bloomer, without whose patient encouragement I would have given up a long time ago.

To my wonderful daughters, Laura and Jessica, who appear to have grown from small girls to young women during the time it has taken to complete this study.

To my husband and my mother for their unending love, encouragement and support.

To Des Johnson whose intervention in my education at primary school, opened up the path that eventually led to my completion of this study.

Finally, this work is dedicated to the memory of my father, who sadly is not here to see its completion.

List of Tables

List of Figures

Chapter One: Introduction

Since the publication of their review of research into assessment and classroom learning, Black and Wiliam (1998a) opened a dialogue on assessment that continues to engage researchers, teachers and policy makers' today. As Broadfoot (2000, p.ix) suggests:

> *Like colonialism before it, the activities associated with educational assessment [...] have steadily advanced during the twentieth century to a point where, at the present time, there can be [...] no mainstream school that is not subject to its sway nor any pupils, teachers or families who do not accept its importance.*

During the first decade of the 21st Century, these debates intensified, and have concentrated on the question of what purpose educational assessment serves. There are those who regard educational assessment as a social practice and a social product that represents "the desire to discipline an irrational social world", and see its primary function as a means of "structuring social hierarchy" (Broadfoot, 2000, pp.ix-x). From this perspective, educational assessment is regarded as a mechanism of social and political control. As Filer and Pollard (2000, p.8) assert:

> *Sociological discourse of assessment presents insights into the fact that, as well as having educational purposes, assessment fulfils a range of political and social functions within modern society. These wider functions are concerned with social differentiation and reproduction, social control and the legitimizing of particular forms of knowledge and culture of socially powerful groups.*

Gipps (1999, p.356) articulates a similar view in suggesting:

The purposes assessment has served in society in the past, as well as the role it plays today, are driven largely by social, political, and economic forces.

However, others suggest that educational assessment can serve multiple purposes including educational improvement, increasing effectiveness of teaching and learning and curriculum reform (Morrison, 1996). Such advocates view assessment as:

... the principal vehicle for advancing the processes of teaching and learning, [and as] increasingly concerned with the improvement of teaching and learning (Gordon, 2008, p.4).

This study is an exploration of the change in teachers' assessment practice in Physical Education (PE) between 2000 and 2005/2006 It is set against the context of the prevailing ideologies and conditions at the time (Broadfoot, 1979). In order to achieve its stated intention, the study will address the following research questions.

Primary research question:

What assessment methods are used in Physical Education at Key Stage 3 in the Riverside Partnership, and how have these developed between 2000 and 2005/2006?

Supplementary research questions:

In what ways do teachers of Physical Education, in the Riverside Partnership, consider the concepts of reliability and validity in their assessment practice at Key Stage 3?

How do teachers of Physical Education, in the Riverside Partnership, make 'best-fit' judgements, as required by National

Curriculum 2000, to decide on end of Key Stage 3 summative attainment levels, which are reported to parents?

Background and policy context for the research

Broadfoot (1979) suggests that assessment is "one of the most political aspects of education", and is directly concerned with issues of "social power and control" (p.122). The background, against which this study is set, was a time of centralisation in terms of national policy for education. The roots of this centralisation can be traced back to the 1980s and it reflects a:

> *...philosophy resting on the belief that it is central government, its ministers and civil servants that must determine not only the shape of the school system but of the curriculum and the methodology of the teaching process. Teachers must therefore be subordinated to a political will based on the notion that only an all-powerful state knows what is best for its citizens (Roy, 1983, p.1).*

The setting and regulating of such political goals in education has an effect on all aspects of teachers' practice in both subtle and profound ways (see Broadfoot, 2000b; Ball, 2003).

In order to contextualise the reality in which the teachers in the case study were working, this chapter briefly sets out the political climate in education at the time of the research (2000 – 2006). Within this broader policy context, it outlines the prevailing accountability culture in education and the accompanying role and power of the Office for Standards in Education (Ofsted) to monitor the implementation of centralised educational policies. It outlines the evolving national interest in assessment between 2000 and 2006, in particular Ofsted (2003b) 'good practice' in assessment in PE and the debates about the nature of knowledge in PE within the framework of the National Curriculum 2000. However, critical reflection on the effectiveness or appropriateness of the prevailing political goals and

associated national education policy development are beyond the scope of this research. Thus, whilst this reality could be critiqued on many levels, including its political stance, its strategy or even its desired educational outcomes, for the purposes of this research it is accepted, as the theatre in which the PE teachers were working. Instead, this study will focus on the changes in teachers' practice in assessment, which occurred, set against this background.

The 1988 Education Reform Act, introduced by the then Conservative government initiated a major transformation in schools in England and Wales that continues today. Much of this reform is beyond the scope of this study. However, the introduction of a statutory National Curriculum and associated assessment arrangements and the provision for an Inspectorate, Ofsted, to police standards at both Local Education Authority (LEA) and school level, initiated the development of a culture of accountability in schools not previously seen anywhere in the UK. The judgements of this national inspectorate, together with the results of pupils' attainment in a range of national assessments were published not only to enhance parental choice but also to identify a hierarchy of schools including a category of 'failing' schools.

Subsequent education policy development in general, and curriculum and assessment policies in particular, by governments from both of the mainstream political parties, have continued to be shaped by what Broadfoot (2000b) and Ball (2003) have defined as a culture of performativity. The performativity discourse is one in which schools and teachers are continually required to improve performance, for example league table position, even if they have already achieved satisfactory standards or grades. By 2000, when the data collection for the present study commenced, the performativity discourse was so pervasive and powerful that teachers and educationalists spoke its language and the performativity technology including its requirement for schools and teachers to plan, teach and assess using standard formats began to dominate practice (Ball, 2003).

Indeed, despite the change of political party in power in 1997, the performativity culture has retained its hold on policy makers and on practitioners, whose performance both collectively and individually is judged in those terms. Reflecting on the first five years of Labour education policies in England, (1997 – 2002) Reynolds (2002, p.97) concludes:

> *[The Labour government] kept in its virtual entirety the 'market-based' educational policies introduced by the Conservative government from 1988 to 1997, involving the systematic tightening of central control on the nature of the curriculum and on assessment outcomes.*

Throughout 13 years of successive New Labour governments to May 2010, government policy, on curriculum and assessment in England, remains extremely committed to the idea that the raising of standards of attainment in schools should be equated with improvement in the grades of successive cohorts of pupils, as they progress through the key stages (1 – 4) of the National Curriculum.

Excellence in Schools (DfEE, 1997), the New Labour government's first major policy paper, set out its intentions to raise standards in education. The overall approach, to implementing subsequent supporting policies, was underpinned by six key principles:

> *Education will be at the heart of the Government*
> *Policies will be designed to benefit everyone*
> *The primary focus is standards*
> *Intervention will be in inverse proportion to success*
> *Zero tolerance for inadequate performance*
> *Government will work in partnership with those committed to raising standards (DfEE, 1997, p.6).*

This "focus on standards" and "zero tolerance for inadequate performance" reinforced a general perception that the most important role for data from National Curriculum assessments was as performance indicators of the standards of attainment achieved by schools. At that time, David Blunkett, the incumbent Secretary of State for Education raised the public profile of such standards in schools, by stating he would resign if the government's national targets based on National Curriculum test data in Maths, Science and English were not met. Simultaneously, the authority of Ofsted, to police the implementation of these policies at school level, was increased.

This discourse about raising standards was dominated by the accountability agenda, where the main purpose of assessment was seen as a way of measuring standards of attainment, rather than as a tool for promoting learning. Alongside this drive for accountability, a number of centrally driven national strategies emerged, with the stated aim of improving attainment, firstly in the primary sector with the literacy strategy and the numeracy strategy from 1998 and1999 respectively. In 2001, in response to central government concerns about the quality of teachers' practice in the secondary sector, the Key Stage 3 strategy was implemented.

The overarching aim of the Key Stage 3 Strategy was to raise pupils' attainment through improving the quality of teaching and learning in schools. Whilst it focused on pedagogy and changing classroom practice, it was an integral part of the government agenda for raising standards in the state education sector. Its implementation, by schools, was subject to periodic inspection by Ofsted. The link between the Strategy and the government's agenda for raising standards was clearly articulated in Ofsted's evaluation of the Strategy in 2005:

> *The Secondary National Strategy, formerly known as the Key Stage 3 Strategy, continues to have a positive influence on pupils' attainment. Since its inception in 2001, the Strategy has made a significant contribution to the steady improvement in the proportion of pupils reaching Level 5 or above in English,*

mathematics and science tests taken by pupils at the end of Year 9 (Ofsted 2005, p1).

The strategy was renamed the Secondary National Strategy for School Improvement (SNS) in April 2005. [For clarity, the abbreviation SNS is used, throughout this thesis, to refer to this overall strategy]. It was extended to include Key Stage 4 and to cover all aspects of teaching and learning, with the assessment strand having been introduced in April 2004. Within the overarching aim of raising standards of attainment, the text discourse for the assessment strand centred on the relationship between assessment and learning, particularly using assessment to improve learning, and thereby improve standards. However, whilst it was underpinned by the research findings of Black and Wiliam (1998a, 1998b) and subsequent work of members of the Assessment Reform Group (ARG), over time it has been mediated by others, such as Ofsted, to drive change in teachers' practice. This interest in teacher assessment at a national level affected the culture for assessment in schools and, more importantly for the present study, in PE departments. Of particular interest to this study is the "Good assessment practice in physical education" published by Ofsted (2003b). A copy is located in Appendix Nine.

This document sets out Ofsted's notion of 'good practice' in assessment in PE and it is substantiated through reference to findings from the previous rounds of school inspections (Ofsted 1995 – 2002). It could be argued that this view of 'good practice' represents a particular approach to assessment in this subject, one with which not all practitioners or researchers might agree. However, this idea of 'good practice' as defined by Ofsted, constituted an authoritative base and one of which teachers were obliged to take notice. Therefore, given the role and power of Ofsted within this prevailing climate of accountability and performativity, and the prevalence of the Secondary National Strategy (SNS) underpinned by the research ideas of theoreticians, it is unsurprising that teachers' assessment practice moved towards this notion of best practice as defined by Ofsted. Detailed reflections on whether this policy was the appropriate policy to achieve

these stated educational aims, or whether the notion of 'good practice' promulgated by Ofsted (2003b) was appropriate, are beyond the remit of this study. However, the changes in teachers' practice, brought about because of the implementation of this strategy are central to this research and are fully considered in Chapter Four.

What should be taught and assessed in PE continues to be controversial. The politics of the formulation of the first version of the National Curriculum for Physical Education 1992 (NCPE, 1992) are well documented (see Evans and Penney, 1995; Bailey, 2005). The determination of the content of the NCPE (1992) took place within a complex and multi-layered context. Part of that context was inter-subject; specifically the long-standing discourse about the relative status of PE within the curriculum and by extension the relative status of PE teachers with their colleagues in other subject specialisms. However, much of the debate was intra-subject and centred on conflicting constructs of PE, underpinned by competing views of PE and sport.

The debate on constructs of PE and its relationship to Sport is examined in Chapter Two of this thesis. However, to set the context in which the teachers were working, it should be acknowledged that defining an agreed concept in PE is problematic. There is much disagreement about what constitutes knowledge in PE amongst PE professionals, including both teachers and theoreticians (see Kirk, 2010; Lee, 2004). Groups differ in their views of the aims of PE, the teaching needed to secure these aims and the means to assess their achievement. The revised NCPE introduced in 2000 demonstrated a marked shift to a more educational perspective of PE from the previous two versions 1992 and 1995, which were dominated by sporting constructs, particularly games. This view of knowledge, represented in the NCPE (2000), formed the basis of what was to be taught and assessed in NCPE (2000) at KS3. It is also relevant in that Ofsted focused their inspections of the assessment practice of the teachers within this conceptualisation of PE. (The detailed Programme of Study for KS3 PE is included in Appendix Five).

As this study does not seek to enter the contested terrain of knowledge constructs in this subject, (Morley, 2008; Lockwood, 2000; Kirk and Gorely, 2000; Green, Smith and Roberts, 2005; Kirk, 2010) the relative merits of this concept of PE are not for debate here. Rather, for the purpose of the present research, it is accepted that this interpretation of PE in NCPE (2000) determined the curriculum and assessment framework in which the teachers were working at the time of the study. However, the impact of this interpretation of PE on teachers' assessment practice, particularly where their personal construct of PE is in conflict with the prevailing conceptualisation of PE determined by NCPE (2000) is an important theme for this study.

In their review of research into assessment and classroom learning, Black and Wiliam (1998a) brought together the research evidence that levels of achievement could be improved by using assessment, not only to summarise learning, but also as part of teaching to help learning and promote pupil autonomy. Their review, "Inside the Black Box" (1998b), and subsequent pamphlet "Assessment for Learning: Beyond the Black Box" (1998c) provided research evidence, in terms of the value of teacher assessment for both summative and formative purposes. Through this research, limitations of external testing were identified and the key role that teacher assessment could play in improving learning, even when used for summative, external purposes was recognised. This view, that assessment can be used to enhance learning rather than merely measure it, was further developed through the work of the Assessment Reform Group (ARG) between 2002 and 2010. It underpins the development of the SNS that has led to the widespread use of assessment for learning in schools today.

As the debates around the advantages and disadvantages of using teacher assessment for summative purposes have evolved, the validity and reliability of using ongoing teacher assessment in this way have become topics of great interest in educational communities (Gipps, Clarke, and McCallum, 1998; ARG, 2002; Harlen, 2005a, 2009; Black et al., 2010). In

examining the evidence concerning the reliability and validity of assessment by teachers used for summative purposes and the extent to which it might be considered dependable, researchers have focused their enquiries on the core subjects of Maths, Science and English.

However, little attention has been paid to these concepts in relation to the subject of PE. That is not to suggest that researchers, academics and policy makers have totally ignored assessment in PE. Indeed, a number of studies of assessment practice in PE do exist. However, they have tended to concentrate on summative assessment practice for examination courses such as GCSE, supporting the implementation of Assessment for Learning specifically in PE, (Casbon and Spackman 2005) as part of the SNS or most recently on supporting the implementation of the Assessing Pupils' Progress (APP) project (Frappwell 2010). In addition, Ofsted has published a number of reviews of inspection evidence specifically in relation to PE practice in schools (1995, 1998, 2002 and 2009). However, contemporary researchers seem to have largely ignored issues of validity and reliability in using ongoing teacher assessment to determine summative attainment levels specifically in PE, at the end of Key Stage 3. This study seeks to explore these issues in an attempt to provide an insight into the changes in teachers' practice between 2000 and 2005/2006.

This study does not concern itself with the arguments of the relative merits of teacher assessment used for either summative or formative purposes. These are articulated in detail elsewhere by other researchers and policymakers (see for example Black and Wiliam, 1998a; Black et al., 2010; Harlen, 2004a, 2005b; Casbon and Spackman, 2005; Mansell, James, and the ARG, 2009; QCA, 2009; Spenceley, 2009). In contrast, the present research is concerned with the way teacher assessment practice is implemented in PE. It starts from the premise that in PE at Key Stage 3, the modus operandi is to use assessment for both purposes, as part of an overall teacher assessment strategy. Furthermore, it accepts the merits of 'good practice' in assessment for both formative and summative purposes, as espoused by other researchers, exemplified by Black et al. (2003) or

Mansell, James and the Assessment Reform group (2009). Finally, this present research into teachers' assessment practice in PE is not seeking to enter the debate of the usefulness or otherwise of teacher assessment as opposed to external testing. In PE, for the purposes of the National Curriculum at least, this argument has been won in favour of teacher assessment.

This introductory chapter seeks to provide a meaningful context for the analysis of the issues and themes of the study. In summation, with the stated aspiration of raising standards of attainment in schools, national policy affected all aspects of state education, from the management and infrastructure of the school system through to learning, teaching and assessment pedagogy and curriculum design. Through a range of education policies, between 1997 and 2010, successive New Labour governments were not only directly shaping what should be taught and how it should be taught but also, more importantly for this research, the ways in which pupils' learning should be assessed.

As previously explained, it is not the purpose of this study to debate these policies, in terms of whether they were the right policies to achieve the stated educational aims, nor is it involved in making judgements about the quality of teachers' practice in relation to the view of knowledge and learning espoused by this policy context (NCPE, 2000). Having accepted at a conceptual level that assessment for both formative and summative purposes plays a part in effective assessment in PE at Key Stage 3, and that assessment contributes in multiple ways to pupils' learning, this study intends to explore the changes in PE teachers' assessment practice, at KS3, set against this background.

Through a case study methodology, which focuses on schools involved in initial teacher education and training, in partnership with Riverside University, the present research explores changes in teachers' assessment practice in PE at Key Stage 3, between 2000 and 2006. Using the work of Harlen (2004a) as a tool for analysis, it will demonstrate that within the

framework of NCPE (2000), at the three data collection points (2000, 2005 and 2006), PE teachers, in the Riverside Partnership, are using an ever-wider range of methods and tools, in order to make dependable assessment judgements at Key Stage 3. It will show that, driven by the prevailing culture of performativity and accountability, PE teachers' assessment practice increasingly moved towards the notion of 'good practice' in assessment in PE, as defined by Ofsted (2003b).

Chapter Two: Literature Review

This chapter provides a critique of a body of literature, which has been selected for its relevance in contextualising and informing the research questions for the present study. There is particular reference to literature extant in the period of the study, which still has relevance. However, later relevant literature is also considered.

Writing in 2010, about the state of teacher assessment in the National Curriculum, Frapwell, the National Subject Lead for PE, suggests:

> *Assessment is perhaps the most difficult area to change behaviours because of the culture of practice that has evolved around the [PE] profession's obsession to convert every bit of progress a learner makes into a number [level] or a grade to create data (p.13).*

In this thesis, which explores change in PE teachers' assessment practice between 2000 and 2005/2006, I have chosen this recent comment, as the starting point for this literature review. For me, as an experienced PE professional, given where we were when I entered the profession (1985), where assessment was an informal reflection on pupils' effort, attitude and behaviour, in largely practical sport and dance activities, it raises the question: How did assessment in PE get to this state by 2010? Why are numerical leveling and accountancy so embedded in the culture of PE teachers' assessment practice, that they are now raised as concerns at a national level?

In order to appreciate the magnitude of these developments, it is important to examine briefly the historical relationship between assessment and PE.

Nature of assessment in PE: Pre-1988

During my own teacher training and education (1983-84), and even in the early stages of my career as a PE teacher in schools, assessment and PE were concepts, which were not considered mutually significant. At this time, little importance was attached to assessment practice in this subject, beyond an informal, reflection on pupils' effort, attitude and behaviour in practical lessons, summarised in an annual report to parents. How PE teachers made these judgements was left entirely to the teacher's discretion, and rarely featured in discussion, even at departmental level. Unlike in those subjects traditionally considered academic, for example Maths, Science and English, the historical nature and purpose of PE in the curriculum, with its physical activity goals and early roots in drill and military preparedness, did not necessitate significant development in assessment practice in PE.

With no National Curriculum, no formal examination judgements to be made and reflection focussing on pupils' attitude, effort and behaviour, rather than on progress and attainment, assessment demands placed on most other curriculum subjects passed by the PE profession. As Carroll (1994, p.2) comments:

> *Assessment debates and reform hardly touched Physical Education. Physical Education teachers were largely left to their own devices [...] in assessment matters.*

This situation remained largely the same until the development of examination courses in PE, in the late 1980s, and the introduction of the National Curriculum in 1992.

Developments in assessment in Physical Education: Post-1988

Since 1988, four major influences directly or indirectly affected the development of assessment practice in PE in schools. These are

- National examination system (GCSE from 1988, A level from 1990)

- National Curriculum (Implemented 1992)
- Ofsted (inaugurated 1992)
- Key Stage 3 National Strategy (from 2002 renamed Secondary National Strategy for school Improvement from 2005 onwards)

Drewett (1991) suggests that the historical lack of importance attached to assessment practice in PE can be accounted for by its lack of status as an examination subject, before 1988. Indeed this increased status for the subject was one of the key arguments put forward by those who pioneered the development of national examinations in PE and related areas (Armstrong and Sparkes, 1991; Carroll 1991, 1994; Kirk and Tinning, 1990).

This increased status of becoming an examination subject, and the subsequent inclusion of PE as a foundation subject in the National Curriculum (1992), meant that PE teachers became more centrally involved in the functions of the school. However, there was also, to some extent, a loss of freedom for the profession to determine its nature and purpose. Like their colleagues in other subject disciplines, PE teachers were now required to engage with the ideologies of assessment, and the developing external and internal accountability agendas, as they affected the wider community of the school. They found they had to account to a variety of audiences including parents, headteachers, local education authorities and governors, in a way that previously had been outside their domain.

Hitherto, in contrast to other major subjects in the curriculum, PE had been characterised by a lack of formal assessment. That is not to say that no assessment took place, however as Carroll (1994, p.19) summarises, the main features of such assessments were:

> *...ephemeral and fleeting evidence, lack of specific criteria, except in award schemes, lack of systematic observation and recording, and reliance on general impressions.*

In response to the requirement for greater accountability, a need to develop approaches to assessment in PE, that were more systematic, was identified (Ofsted, 1995; Mawer, 1995; Piotrowski and Capel, 2000). With little formal training in assessment up to this point, and limited experience of assessment, PE teachers were required to establish valid methods of assessing pupils' progress against a number of examination syllabi. As a result, mirroring the practice of traditionally more academic subjects, the assessment strategies at CSE, then GCSE and A level, tended towards an academicism of the subject, relying primarily on cognitive knowledge, skills and understanding that could be tested through formal exams and written assignments, supplemented with teacher observation and assessment of practical skills.

This development of assessment practice, for the new examination courses, including the separation between practical and cognitive goals, began to affect approaches to assessment of practical aspects of the subject in non-examination classes. In turn, it began to influence teachers' views of the nature and purpose of PE. For many practitioners there was a shift from the 'process' of engaging in PE as one of the primary goals for the subject to an emphasis on the 'product' of performance. McChonachie-Smith (1991) identified a tension between what is defined as capability in PE and what it is possible to assess. To clarify, a 'product' of PE performance, such as a well executed vault in gymnastics, is more easily assessed through observation than the 'process' of engaging in PE, for example learning how to evaluate either one's own or others' work in order to improve future performance. As Piotrowski and Capel (2000, p.108) caution:

> *Care is required to ensure that the content of Physical Education is not distorted to accommodate that which is amenable to measurement at the expense of equally valuable but less easily assessed components.*

The drive for greater accountability, which underpinned the introduction of the National Curriculum (1992) and its assessment, was also evident in the power of Ofsted to inspect schools' compliance with centralised policies. The impact of the climate of accountability, that was created by the terms of the Education Reform Act 1988, was far reaching. As part of this accountability agenda, schools needed to be seen to do well in the inspection process and to comply with criteria for 'good practice', as defined by Ofsted. This has proved problematic in PE, where practice in this subject has been repeatedly criticised by Ofsted, (1995, 1998, 2002, 2009). In 1995, Ofsted (1995, p.5) reported that:

> *...the quality of assessment reporting and recording needs to be improved at all key stages. Teaching should be informed by results of assessment.*

In its summary report of Initial Teacher Training subject inspections (1996 - 1998), Ofsted (1998) again concluded that assessment within PE was problematic. This, it suggested, was directly linked to the lack of good models of assessment practice, within many PE departments in schools to help trainees develop their practice. In the face of such criticism, with a lack of training, experience and models of good assessment practice, and given the power of Ofsted to determine the criteria against which teachers' practice is measured, a tendency to move towards more formal approaches to assessment in PE, is unsurprising.

Satterly (1981, p.352) suggests that formal assessment can be defined as:

> *...assessment conducted in situations solely for that purpose, whereas informal assessment is assessment conducted while pupils are carrying on normal classroom activities.*

Given the transitory nature of practical performance in PE, reaching accurate judgements of pupils' attainment is notoriously difficult (Ofsted, 1995, 1998, 2009). The difficulties are compounded when the assessment is

informal in nature, and undertaken in the context of learning and teaching in classes of 30 or more pupils (Piotrowski and Capel, 2000). Moving towards more formalised approaches to assessment allowed teachers time to judge pupils' attainments more accurately against clearly specified criteria. Satterly's (1981, p.352) view that:

> *...the more formal the mode of assessment, the more the assessment process itself is open to scrutiny.*

gained credence during this move to more formalised approaches to assessment. Piotrowski and Capel (2000, p.107) linked more formal approaches to increased integrity in assessment, suggesting that:

> *...more formal methods can help guard against a lack of fairness.*

In 2003, drawing from the evidence gathered from inspections since 1995 and the specific concerns relating to assessment practice in PE, Ofsted published a set of recommendations to disseminate their notion of 'good practice' in assessment in PE nationally. This document was a drawing together of 'good practice' seen in school PE departments, as perceived by the PE Ofsted inspectors, and was accompanied by a number of national dissemination events, with a stated aim of leading to improvement in teachers' assessment practice over time.

At the time, this was seen by the PE profession as a welcome attempt by Ofsted to move from simply identifying and criticising poor practice (Ofsted 1995, 1998, 2002) to identifying and disseminating their notion of "good" practice. The fact that these recommendations were formulated based on Ofsted inspection evidence, added credence and weight to their status. The lack of criticality, with which this was received, by both teachers, educationalists and researchers alike, reflects the prevailing culture of performativity and accountability in education. There was a broad consensus that rather than trying to second guess what Ofsted was looking

for in terms of assessment practice in PE, the profession had now been given the answers. Rather than critique this conceptualisation of 'best practice', the profession almost unanimously agreed that its task was to develop assessment practice in PE, in line with this notion. This perspective influenced both teachers in schools, and those engaged in teacher education.

In reflecting on whether PE teachers really are engaging in 'good practice', in using the Ofsted (2003b) principles, it could be argued that rather than representing 'good practice', the approach to assessment, promoted by Ofsted (2003b) actually resulted in negative consequences, which are still issues of concern today. One such example would be the focus on using levels and sub-levels:

> *Levels are recorded using +/- to indicate subtle differences between pupils (p.5).*

There is no requirement in the NCPE (2000) to use levels other than at the end of Key Stage 3. However, influenced by the assessment practice of teachers in the core subjects such as Maths and Science, and in an attempt to meet the Ofsted criteria, the inappropriate use of levels and sublevels has become an issue of concern. In short, it lead to PE teachers teaching to the levels, rather than them being used to support learning. This impact on PE teachers' assessment practice is still an issue in 2010, and is summarised by Frapwell (2010, p.17):

> *The frequent testing of levels achieved and in focusing on [pupils'] deficiencies in order to reach the next level, [...] the levels have become the goals rather than enhancing learning or the processes by which to progress.*

Even at the time of writing (2011), trainee teachers within the Riverside Partnership frequently report that they have been encouraged to 'level' pupils in every lesson. It should be noted that this use of levels could be an

unintended outcome of the way teachers' practice evolved in response to the Ofsted (2003b) principles. However, this example serves to illustrate the power of Ofsted to impact on teachers' practice. It is clear that the authoritative basis of these Ofsted (2003b) principles has so influenced practice that they continue to affect assessment in PE today. Frapwell (2010, p.14) concludes:

> *...even though QCA (1999) issued guidance around the use of levels and assessment linked to the NCPE (2000), much of the profession have completely ignored this guidance and used levels in a way that was never intended.*

The performativity climate in schools similarly affects teacher education institutions. There is a direct link between the allocation of teacher training places to institutions and their Ofsted inspection outcomes. Within the framework for inspecting initial teacher education, at the time when this study took place, courses were graded according to three strands, one of which was assessment. The dependence of Higher Education institutions, on success in such inspections to determine the amount of funding, and therefore the viability of individual courses, at a time of contracting numbers in PE, resulted in an unprecedented drive to change assessment practice in line with Ofsted's notion of 'good practice'. Similar to many other institutions, the curriculum at Riverside University was revised in order to develop the trainee teachers' assessment practice in line with these Ofsted recommendations. It is interesting to note that, despite these developments, whilst there is evidence of some improvements, when measured against the Ofsted criteria, inspection evidence continues to suggest that teachers' assessment practice in PE is still considered problematic. According to Ofsted (2009, p.3):

> *The better schools visited assessed, recorded and tracked pupils' progress systematically. However, inconsistencies remained in judging pupils' standards and achievements accurately. Most of the secondary schools visited did not assess*

students' standards and achievement in core physical education at Key Stage 4.

Whilst it would be possible to debate further this notion of 'good practice', for the purposes of this study, it is accepted as the prevailing educational standard, against which PE teachers' assessment practice was measured. How teachers' practice changes towards this notion of 'good practice', driven by Ofsted is a key theme for the present study.

The fourth influence on the development of assessment practice since 1998, identified earlier in this chapter, is the Key Stage 3 National Strategy, later renamed Secondary National Strategy. For the purposes of clarity, this section outlines the main aims of the SNS, and then a critique of the strategy is offered.

The SNS, piloted in September 2000 and rolled out across all state schools in England and Wales in 2002, is a government driven strategy for whole school improvement. At its core are the four key principles of:

Expectations: Establishing high expectations for all pupils and setting challenging targets for them to achieve
Progression: Strengthening the transition from Key Stage 2 to Key Stage 3 and ensuring progression in teaching and learning across Key Stage 3
Engagement: Promoting approaches to teaching and learning that engage and motivate pupils and demand their active participation
Transformation: Strengthening teaching and learning through a programme of professional development and practical support (DFES, 2000, p.3).

Whilst this strategy eventually covered all aspects of schooling, from behaviour and attendance to school leadership, its primary aim was to raise

standards of pupils' attainment through improving the quality of teaching and learning in schools. This strategy was very significant in affecting teachers' practice, as not only did it espouse aspirations and principles, but also it was underpinned by a programme of professional development. Targeted at a whole school level, including head teachers, senior managers and school governors, the strategy set out to achieve whole school improvement and raise standards of attainment in the state sector of education in England and Wales. Beginning with Key Stage 3, and a particular emphasis on Maths and English, the strategy was subsequently broadened to include all subjects across the secondary age phase and was renamed The Secondary National Strategy, (SNS) in 2005.

Significant resource has been, and continues to be set aside, in order to develop teachers' practice in line with its stated principles. In addition, Ofsted has evaluated the progress of the Strategy from the very first year of the pilot, and over time, the Ofsted framework for inspections has evolved to take account of the extent to which schools are developing in line with these principles.

In setting out the four main principles, a number of foci were identified. For the purposes of the present research, the focus on assessment from 2004 onwards, is of particular relevance. This nationally driven, well-funded strand of the SNS was underpinned by research evidence, in relation to the value of teacher assessment, with particular emphasis on the work of Black and Wiliam (1998c) and the ARG (1999, 2002, 2006). Its purpose was not only to raise awareness of assessment issues in schools, but also to develop teachers' practice in assessment, through a policy of whole school continuing professional development (CPD).

In 2004, six years on from Black and Wiliam's review of the research evidence, in relation to the benefits of teacher assessment in improving pupils' learning, where they suggested that:

> *The improvement of formative assessment cannot be a simple matter. There is no quick fix that can be added to existing practice with promise of rapid reward. [...] This can only happen slowly and through sustained programmes of professional development and support...for lasting and fundamental improvements in teaching and learning can only happen in this way (Black and Wiliam, 1998b, p.15).*

there was a real commitment at national level to realising this radical overhaul in teachers' assessment practice.

Having briefly outlined the principles and focus of the SNS, this section offers some reflection of its role in contributing to the prevailing performativity culture in the schools at the time of the present study. Given that the work of Black and Wiliam, (1998a) and ARG (1999, 2002) are cited in the SNS documentation, the question arises as to whether there is a tension between a research-informed view of teaching and learning, and its use by the state to drive change in teachers' practice. I might argue that is simply an excellent example of research informed teaching. On the other hand, it could be perceived as the centralised application of political power to control the autonomy of teachers, in order to manipulate them to be compliant instruments of the state. The fact that this strategy draws upon a significant body of research undertaken by these leading theoreticians of the day, should not inhibit the questioning of this attempt to change teachers' practice, on such an unprecedented wholesale scale.

One interpretation might be that this was a rare time, when what is considered 'good practice' by educational theorists and authenticated by high quality research, coincides with the political will of a government: where appropriate funding, CPD and structures for implementation are made available, and as such it reflects a real commitment to improve the educational experience of all pupils. An alternative view might be that this is a further example of the centralisation of national education policy, with the State using the research evidence, to justify driving change in teachers'

practice in a particular way, in order to meet specific targets in raising standards of attainment in schools.

Whilst the party in power at the time of this study, Labour, is not the same as at the time when Ball (1994) was writing, Conservative, his comments about the relationship between teachers, the State and the role of teaching still have resonance for this research. Discussing the imposition of a National Curriculum and direct and indirect interventions into pedagogical decision-making, he suggests:

> *...there is an increase in the technical elements of teachers' work and a reduction in the professional. Significant parts of teachers' practice are now codified in terms Attainment Targets and Programmes of Study, and measured in terms of Standard Assessment Tasks. The spaces for professional autonomy and judgement are (further) reduced. A standardisation and normalisation of classroom practice is being attempted (Ball, 1994, p.49).*

Whilst one interpretation of the SNS might be as a research led CPD programme of support to facilitate improvement in teachers' practice, an alternative interpretation might be the imposition of a central approach to teaching and learning that reduces teachers' professionalism and autonomy in their decision making. Taking this stance, it follows that the SNS reflected a centralised view of teaching and learning, in which edicts from political leaders of the day, drawing on the research of leading theoreticians of the day, promotes a particular view or approach to learning, teaching and assessment. From this perspective, the SNS is seen as a move to change teachers from creative, autonomous, reflective practitioners to apprentices of a craft, which can be learned, with suitable training, by anyone. In contrast rather than a centralised determination of a particular view of teaching, it might be argued that these moves to standardisation and conformity are positive in raising the educational experience of all pupils.

No matter how the intention is interpreted, one of the expected outcomes was change in teachers' practice, in line with the principles and approaches promoted by the SNS. The extent to which such change occurred was reported on, by Ofsted, in each of its evaluations of the SNS, and indeed the approaches and techniques promoted through the strategy, became part of the success criteria for Ofsted inspections in all subjects.

In considering the evidence of the promotion of a Strategy approach to teaching and learning, the key findings from Ofsted evaluation for 5th year of SNS are of interest. Ofsted (2005, pp.4-5) reported their key findings:

> *As a result of Strategy guidance, departmental schemes of work are now better structured and more comprehensive, although it is still unusual for them to provide guidance on teaching pupils of different abilities.*
> *The quality of teaching and learning continue to improve as teachers apply Strategy techniques. The best lessons include a wide range of teaching strategies, with more emphasis on pupils thinking for themselves.*
> *In less effective lessons, teachers often use recommended structures and approaches too mechanistically with too much emphasis on content rather than developing conceptual understanding.*
> *The use of assessment for learning is good in only a few of the schools and unsatisfactory in a quarter, but Strategy support for this is still at an early stage.*
> *In about a tenth of the schools, there is still a lack of commitment to the Strategy amongst teachers, mainly those who have little knowledge of its potential.*

The use of the word "recommended" in point 3 is interesting and worthy of note. The link between not using the Strategy techniques appropriately and ineffective lessons indicates clearly the view that there is a Strategy way to teach and that those not using it appropriately are judged ineffective. Point 5

is also worthy of consideration. Whilst it might well be that those resisting adopting the Strategy way of teaching are doing it from a perspective of ignorance of "its potential", it might also be that these schools are content with the effectiveness of their own practice, for their pupils and community. Nevertheless, the stating of this in these terms, to suggest that teachers would only not adopt it if they were in some way deficient in their ability to understand that it is good for them, could be interpreted as undermining of the professionalism of such schools and their leadership.

It is acknowledged that the stated aim of the SNS was to raise standards in teaching and learning. Therefore it is unsurprising that a particular view is being promoted. Given the level of funding invested and the power of Ofsted to measure success against the SNS and the contribution of those inspection outcomes to judge overall success in schools, it would be unsurprising in this performativity and accountability climate, for teachers' practice to move towards a Strategy notion of 'good practice'.

Further discussion of whether this Strategy notion of 'good practice', as with the Ofsted (2003b) notion of 'good practice' in assessment in PE, represents 'good practice', are beyond the remit of this thesis. However, the SNS is important in the present research, as it contributes to the culture in the schools in which the participants were working, at the time of the study, As a result, it may have contributed to changes in their assessment practice.

Exploring such change, in the policy context that was the background to this study, is a major focus of this research. In examining the changes in the assessment practice of the PE teachers in Riverside Partnership, lessons about the wider influence of political agendas in education may be learned. How change occurs could have particular resonance for teachers and trainee teachers entering the profession in 2011, given the election in May 2010 of a conservative led coalition government. Already, the new Secretary of State for Education, Michael Gove, has indicated that many of the previous New Labour government's key policies for education will be replaced. This includes provision for a much-reduced National Curriculum from 2013, in

which it is suggested that PE may not even be included, and a change to the role for Higher Education Institutions in Initial Teacher Training and Education is proposed.

Whilst the educational politics of New Labour and Conservatism are not for debate in this thesis, the power of the state to affect teacher's classroom practice, regardless of whichever political party is in government, should be recognised. Ball (1994, p.50) cautioned:

> *... significant changes in teachers' classroom practice can now be achieved by decisions taken at a distance about assessment regimes or curriculum organisation.*

As both Broadfoot (2000b) and Ball (2003) have suggested, the setting and regulating of political goals in education has an effect on all aspects of teachers' practice in both subtle and profound ways and the impact of such wider political change on teachers' classroom practice is of interest to this study. Having examined the main influences on the development of assessment practice in PE since 1988, I will now consider the national developments in assessment practice, which have relevance for the present research.

National developments in assessment practice

A major educational aim, in the first decade of the new millennium, (2000 – 2010) has been to promote assessment as an integral part of the learning and teaching process: assessment is seen to be a tool of the curriculum, and significant debate about its purposes has occupied theoreticians, policymakers and teachers. Whereas, an interpretation of assessment as a measurement of learning (summative) dominated the discourse at the end of the twentieth century, since the work of Black and Wiliam (1998a, 1998b, 1998c), and the subsequent development of their research ideas by the ARG (1999 to 2010), assessment is now also interpreted as a tool to promote learning (formative). Because of the SNS, the term assessment for learning

is equally part of the vocabulary of both practitioners and theoreticians alike.

The merits of the argument to entrust assessment for both formative and summative purposes to teachers, continue to be vigorously debated in the research literature (Black and Wiliam, 1998a and 1998b; Harlen, 2004a and 2005a, 2005b; Black, 2005; Stobart, 2008; Mansell, James and the ARG, 2009; Newton, 2010). It is argued that there would be an achievement of synchronisation in assessment practice, if teachers were responsible for both summative and formative assessment. In this context it is argued that summative assessments, including so called 'high stakes' assessments, for example those which impact upon pupil career choice and progression, such as GCSE or A levels, can be informed through the formative approaches used by teachers in assessing pupils' ongoing progress and attainment. For the advocates of such an approach, a primary advantage suggested is that if teacher assessment held the responsibility for both formative and summative purposes, then a truer, more rounded assessment of pupil learning and attainment can be reached.

As these debates about the usefulness of teacher assessment for both summative and formative purposes have evolved, so too has an interest in the concepts of validity, reliability and dependability in relation to teacher assessment practice. Implicit in this notion of assessment being a tool of the curriculum is that the assessment, particularly when used for summative purposes, should still be expected to be valid and reliable. This educational aim has resulted in more than a decade of research, debate, development and re-conceptualization of the issues involved in teacher assessment and its dependability.

In this study into changes in PE teachers' assessment practice between 2000 – 2005 / 2006, the ways in which teachers, in the Riverside Partnership paid attention to these concepts, in relation to their changing assessment practice, is considered. Therefore, in this section, I will briefly examine the complex

concepts of validity and reliability in relation to assessment and clarify the definitions of each that are accepted for the purposes of the present research.

Regardless of the purpose of assessment, summative or formative, concepts of validity and reliability are complex and not unproblematic. How these are understood depends very much, on how learning and knowledge are understood. From a modernist perspective in which the "notion of true knowledge can be seen as a mirror of reality", (Kvale, 1995, p.1) and where knowledge is understood as external and objective, then validity can be determined as the extent to which an assessment is an accurate reflection of such an objective truth. However, in other interpretations of knowledge, where knowledge is seen as subjective and constructed by the learner and their engagement with the social world then validity as a concept is more problematic to define. Kvale (1995, p.1) suggests that from a post-modern perspective:

> *...the concept of an objective reality to validate knowledge against has been discarded.*

In an examination of different notions of validity, Winter (2000) suggests that there is no single fixed or universal interpretation of this concept and it cannot be explored in isolation from notions of truth. Drawing on Foucault's (1974) work on the nature of truth and its multiplicity, he concludes that for different truths different approaches to validation are required. He suggests that rather than interpreting validity as the extent to which the assessment is measuring what it intended to measure, it would be more enlightening to ask, "Is it measuring the kind of "truth " it intended to measure" (p.10).

Validity and reliability are important in my study from two perspectives. The first is in relation to the focus of the study: changes in PE teachers' assessment practice between 2000 and 2006. The second relates to the methodology of the study itself. In an attempt to make sense of these concepts for the present research, the interpretations of Easterby-Smith, et al. (2002) Table 2.1 were helpful.

Table 2.1 Interpretation of notions of validity and reliability		
	Positivist view point	Phenomenological view point
Validity	Do the measures correspond closely to reality?	Have a sufficient number of perspectives been included?
Reliability	Will the measure yield the same results on different occasions (assuming no real change in what is to be measured)?	Will similar observations be made by different researches on different occasions?

(Adapted from: Easterby-Smith et al., 2002).

In considering the issues of validity and reliability in terms of the methodology for the case study itself, I accepted the interpretation of these notions from the phenomenological standpoint, as defined by Easterby-Smith et al. (2002). Based on the notions of credibility and transferability (Guba and Lincoln, 1995 and Guba, 2005), the key questions that need to be asked in judging the quality of my work are:

> *How credible are the particular findings of the study? [...] How transferable and applicable are these findings to another setting or group of people? (Marshall and Rossman, 2006, p.201).*

Notions of transferability as they relate to my study are further explored in Chapter Three of this thesis.

In seeking to clarify my understanding of these concepts in relation to PE teachers' assessment practice, my starting point reflected Easterby-Smith et al. (2002) positivist viewpoint. My working definitions at the time of conceiving the research were as follows:

> *Validity is the extent to which the assessment is assessing what it claims to be assessing, and reliability is interpreted as the extent to which the same results would be found on other occasions or by other assessors.*

The work of Harlen (2004a) has been very influential in developing my conceptual framework for this research. The importance of this work, in shaping the definitions of validity and reliability that were accepted for the focus of my study, is now examined.

In 2004, Harlen led a systematic review of the research studies, available at that time, which were concerned with the reliability and validity of using teacher assessment for summative purposes. The proposal for this review resulted from the work of the Assessment Reform Group (ARG) over several years (1999 – 2004) about the usefulness of teacher assessment in both summarising and informing learning. It is clear from the work of the ARG (1999 – 2010) that assessment by teachers has the capability to provide summative information about learners' achievement, particularly as teachers can take into consideration pupils' performance across a full range of learning activities.

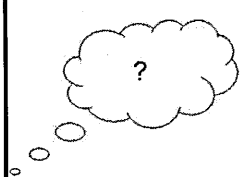
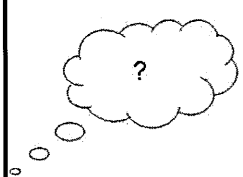
In an earlier review into the impact of summative assessment and tests on students' motivation for learning, Harlen and Deakin Crick (2002) concluded that whilst high stakes tests had a de-motivating effect on the pupils' learning, summative assessment judgements are a necessity in providing information about pupil progress and attainment to a variety of stake holders, including teachers, parents and the pupils themselves. It also concluded that to be effective, summative assessments should interfere as little as possible with the pupils' learning process and should address the full range of learning outcomes within the given curriculum. These similarities with best practice from formative teacher assessment practices were noteworthy. However, whilst in several countries, assessment by teachers has been adopted as the prime source of information in national and state assessment systems, nevertheless in other countries it is considered

"unreliable and subject to bias" (Harlen, 2004a, p.1). These debates, in England in 2004, were less developed than they are today, 2010. Harlen (2004a) sought to test this assumption by examining the available research evidence about the dependability of summative assessment by teachers and the conditions that affect it. This review sought to answer the following key questions:

> *What is the research evidence of the reliability and validity of assessment by teachers for the purpose of summative assessment?*
> *What conditions affect the reliability and validity of teachers' summative assessment? (p.1)*

In Table 2.2, teacher assessment is contextualised, in relation to the more traditional summative practices of external tests and exams, in terms of its potential for classroom impact. It locates the focus of the review, which is to seek evidence to inform the two empty boxes.

Table 2.2: Validity, reliability and classroom impact

	Validity - does the approach give a fair assessment of what it claims to measure?	*Reliability* - are the outcomes of the assessment reproducible?	*Classroom impact* - what impact does this assessment have on the classroom?
External tests / exams	External tests and examinations are *perceived* as having high levels of validity. However, the skills and knowledge being tested do not always appear to be transferable, and the tests can be viewed as artificial rather than authentic. The claim of high validity is not well supported by evidence.	External tests and examinations *are perceived* as having high levels of reliability. Despite the use of rigorous mark-schemes, moderation and scrutiny procedures, the claim of high reliability is not well supported by evidence.	External tests and examinations *are known* to have *negative impact on students' motivation* for learning, *negative impact on curriculum content* ('what is taught is what is tested'), and *negative impact on teaching* approaches (excessive test practice, and 'chalk and talk' approaches predominate)
Teacher assessment			*Teacher assessment*, used for formative purposes, *benefits teaching* (through a greater emphasis on responding to students' known needs), *benefits learning* (by encouraging activities that promote understanding), and *raises standards* of student performance.

(Source: EPPI-Centre, 2006).

Before detailing the relevance of this work to the present study, a brief reflection on the methodology for this review is now presented. Systematic reviews emerged out of a movement to use research evidence to inform both policy making and practice.

> *Systematic reviews aim to find as much as possible of the research relevant to the particular research questions, and use explicit methods to identify what can reliably be said on the basis of these studies (EPPI-Centre, 2011).*

The project that evolved into the Evidence for Policy and Practice Information Centre (EPPI-Centre) was established in 1992. Since 2000, the brief of the EPPI-Centre has expanded through funding from Department for Education and Skills (DfES) to support groups wishing to undertake reviews in the field of education. The EPPI-Centre (2011) claims that the key features of a systematic review are that:

> *Explicit and transparent methods are used*
> *It is a piece of research following a standard set of stages*
> *It is accountable, replicable and updateable*
> *There is a requirement of user involvement to ensure reports are relevant and useful.*

The methodology for the review followed the procedures devised by the EPPI-Centre with a wide-ranging search for published research studies that dealt with some form of summative assessment conducted by teachers, involving pupils in school, aged between 4 and 18. A total of 431 studies was located. However, after exclusions, for a variety of reasons, only 30 were included in the in-depth review. A detailed summary of the methods used, including the systematic map of the review is located in Appendix Twelve.

As can be seen from the systematic map, there were no studies that looked at assessment in PE, and the majority of studies focused on Maths. There were 13 studies that concentrated on the Secondary sector, with a further 6 that considered both Secondary and Primary. 15 of the studies were from England. The most common purpose of the assessment in the studies was for national or statewide assessment programmes, with 6 studies related to certification and another six to informing parents. As might be expected in the context of summative assessment, most focused on teachers' use of external criteria. There was limited research on student self-assessment or teachers using their own criteria.

This systematic map of the studies included is important when considering the findings of the review. Whilst there are similarities between the studies in the review and the contexts in which PE teachers undertake summative assessment, there are also significant differences. For example, the assessment at KS3 NCPE (2000), which is the focus of the present study, does involve external criteria, insofar as the standards are defined by the End of Key Stage Level Descriptions. However, teachers devise and use their own criteria in order to reach their judgements on pupils' attainment. Notwithstanding such differences, I felt that the Implications for Practice, proposed by the review, could provide a useful framework for exploring changes in the PE teachers' assessment practice in Riverside Partnership. Given that the review had focused on other curriculum subjects, I was particularly interested in examining how these conditions that affect dependability in teacher assessment, when used for summative purposes, might be in evidence in the PE teachers' practice.

I combined the key findings of this review, specifically the Implications for Practice, with the Ofsted (2003b) key principles of "Good assessment practice in PE", into an organising framework for analysing the data collected for the present study. My decision to combine these two was grounded in the view that between them, they took into consideration the most recent research evidence and the latest inspection evidence of PE teachers' assessment practice, at the time of the study (2000-2006). In constructing this framework in this way, I also sought to apply research on assessment, undertaken more broadly in education to an understanding of assessment practices in physical education.

This framework was devised as a tool of analysis to help gain an understanding of the changes in teachers' assessment practice in PE at KS3 during the study period. Whilst this was useful as a lens through which to explore such change, the question as to whether the practice identified in the framework is good or otherwise for PE is beyond the remit of this thesis. There is no assumption that a close affiliation between the framework and PE teachers' assessment represents 'good practice'. Rather, it will be used to

illustrate the ways in which PE teachers' assessment practice changed within the prevailing policy context at the three data collection points for the study (2000, 2005 and 2006).

Whilst validity, reliability and dependability are complex concepts, and link to how learning and knowledge is understood, for the purposes of this review, Harlen (2004a) appears to treat them in an almost entirely unproblematic way. Whilst there is some acknowledgement that "different forms of validity derive from different ways of estimating it" and "construct validity is an useful overarching concept", there is no further theorising of these concepts in relation to knowledge. For the purposes of this review, Harlen (2004a, p.7) accepts them as:

> *Reliability refers to how accurate the assessment is (as a measurement); that is, if repeated, how far the second result would agree with the first.*
> *Validity refers to how well what is assessed, matches what it is intended to assess.*

In addition, Harlen (2004a, p.7) suggested that:

> *Since reliability and validity are not independent of each other - and increasing one tends to decrease the other - it is useful in some contexts to refer to dependability as a combination of the two.*

Adopting a similar approach, the definitions for validity, reliability and dependability offered by Harlen (2004a) were accepted for this exploration of change in PE teachers' assessment practice. These are consistent with the definitions subsequently offered by Mansell, James and the ARG (2009, p.12):

> *Reliability and validity are central in all types of summative assessment made by teachers.*

> *Reliability is about the extent to which an assessment can be trusted to give consistent information on a pupil's progress; validity is about whether the assessment measures all that it might be felt important to measure.*

Like Harlen (2004a) Mansell, James and the ARG (2009, p.12) regard dependability as a combination of reliability and validity:

> *Together maximum validity and optimal reliability contribute to the dependability of assessments – the confidence that can be placed in them.*

This work was very influential in the development of the present research. The data collected for the present research was reviewed to examine the extent to which the conditions for the dependability of formative assessment used for summative purposes, as identified by Harlen (2004a) were in evidence in the practice of the PE teachers in the schools in the Riverside Partnership. The key findings of this review for teachers' assessment practice, which are relevant for the present study and their interpretation for the present research are summarised in Table 2.3 below. A full version of the conclusions and implications for research, policy and practice are located in Appendix Three.

Table 2.3 Key Findings from Harlen (2004a):

Finding from Harlen (2004a) review	Interpretation for present research
Teachers should not judge the accuracy of their assessments by how far they correspond with test results, but by how far they reflect the learning goals.	Accuracy of assessment judged by extent to which they reflect learning goals
There should be wider recognition that clarity about learning goals is needed for dependable assessment by teachers.	Clarity in learning goals increases dependability of assessment
Schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers are sustained: for example, by protecting time for planning assessment, in-school moderation.	Whole school commitment to providing time for in-school moderation, planning
Schools should develop an 'assessment culture' in which assessment is discussed constructively and positively, and not seen as a necessary chore (or evil).	Assessment culture discussion of assessment in a positive climate

Having reviewed the evidence in relation to the dependability of teacher assessment used for summative purposes, Harlen (2005) led a further systematic review to examine the research evidence of a variety of different claims and experiences, related to the impact, on pupils and teachers, of the use of teacher assessment for summative purposes.

This review of research evidence substantiated claims that through using teacher assessment for summative purposes, teachers can reach judgements in relation to the whole profile of their pupils' achievement and that they are less threatening to pupils. Therefore, they give a truer account of all aspects of their learning and progress. This review also concluded that assessment by teachers allows for more appropriate learning strategies to be used, which allow for each pupil to best achieve their full potential and that ongoing teacher assessment can be used to help learning as well as to summarise achievement. Thus, the case for teacher assessment was further strengthened in the research evidence, and there was a mechanism to change practice in schools through the developing SNS. This is particularly relevant to the present study, where assessment practice in PE is considered against a background of change in assessment culture at national level.

Purposes of Assessment: Development of AfL and AoL

Evidence that raised levels of achievement result from using assessment in a different way, as part of teaching to help learning was brought together by Black and Wiliam (1998a) in their review of research on classroom assessment. This review identified the limitations of external testing and identified the key role that teacher assessment could play in improving learning even when used for summative purposes. This work, and that which followed, through the auspices of the ARG, of which Black and Wiliam were founder members, has been very significant in shaping the reforms in assessment policy and practice in England and Wales.

Definitions of each approach have received considerable attention in the literature in recent years (Black and Wiliam, 1998a; Black and Wiliam, 2009; ARG, 2002; Black et al., 2003; Harlen, 2005a; Gardner, 2006; Black et al., 2010). Whilst there may be slight variation in the detail of each author's definition, for the purposes of the present research, they can be summarised as follows; Assessment for learning (AfL) or formative assessment is defined as assessment that is used to inform or promote learning. Assessment of learning (AoL) or summative assessment is assessment that sums up learning at a given point. Harlen (2005a) distinguishes between these two main purposes of assessment; the former is defined in terms of its role in helping learning and the latter in terms of its role in summarizing learning. From this interpretation, it is the purpose of the assessment that defines whether it is formative or summative, not the process by which it is undertaken.

Following Black and Wiliam's (1998a) influential review of research on learning and assessment, and their promotion through the SNS (2004) these two purposes, traditionally termed formative and summative assessment, are now also commonly termed 'assessment for learning' and 'assessment of learning.' The terminology 'assessment for and assessment of learning' has succeeded in locating the importance of assessment within the learning cycle rather than as a bolt on summative activity (Winter 2003). For this

study, it is therefore important to examine both the differences between these purposes and their interrelationships. For whilst the focus of the present research is on the reporting of learning at Key Stage 3 in PE, (AoL) the process of gathering evidence to reach these judgements is frequently assessment for learning (AfL).

In attempting initially to raise awareness and subsequently to change practice, the Qualifications and Curriculum Authority (QCA) summarises formative assessment as being concerned with 'assessment *for* learning' and summative assessment being concerned with 'assessment *of* learning.' These summaries that are based on the work of Black and Wiliam (1998a, 1998b) and the ARG (1999) are significant in that they have been widely disseminated to schools and form the basis of many teachers' understanding of these concepts:

> *Central to formative assessment or 'assessment for learning' is that it is embedded in the teaching and learning process of which it is an essential part; shares learning goals with pupils; helps pupils to know and to recognise the standards to aim for; provides feedback which leads pupils to identify what they should do next to improve; has a commitment that every pupil can improve; involves both teacher and pupils reviewing and reflecting on pupils'' performance and progress and involves pupils in self-assessment.*
> *(QCA, 2001, p.7).*

The emphasis in the definition is on the on-going formative nature of such assessment and that it involves the learner in the assessment process with the goal of increasing pupil autonomy in their learning. This was a very significant shift in pedagogy for most teachers, who, even in continuous assessment activities, regarded assessment as primarily the responsibility of the teacher. This has raised the profile of peer approaches to assessment and more significantly self-assessment by pupils.

In 2006, Marshall and Drummond undertook a study to explore the ways in which teachers enact Assessment for Learning (AfL) practices in their classrooms. Their starting hypothesis was that:

> *AfL is built on an underlying pedagogic principle that foregrounds the promotion of pupil autonomy (Marshall and Drummond, 2006, p.133).*

Through lesson observations and teacher interviews, they examine the difference between "the letter" and "the spirit" of AfL. From their study, teachers, whose lessons encapsulate "the spirit" of AfL are characterised by a belief that its value is not only to promote learning but more crucially to promote pupil autonomy. They found evidence of this characteristic in only 20% of the lessons observed. However, there was wider evidence of teachers' practice conforming to the "letter" of AfL. According to Marshall and Drummond (2006), this is identified as using the tools or approaches that help teachers to improve their practice in using assessment to promote learning with their pupils. However, such practice lacks a clear commitment to the underpinning pedagogic principle of developing learner autonomy. Thus in this model, teachers may be said to be "doing" AfL. For example they share learning objectives with their pupils; adopt a range of teaching and assessment strategies, including for example the use of peer or self-assessment. However what distinguishes between the "letter" and the "spirit" of AfL is their understanding and acceptance of their role in promoting pupil autonomy.

Reflecting back on his experience of 10 years involvement in AfL, Spenceley (2009), one of the Science teachers involved in the Kings-Medway - Oxfordshire Formative Assessment Project (KMOFAP) confirms the importance of this view:

> *The key message all along was that AfL was seen to be as a style of teaching, rather than something to add to an already manic*

workload. Not more planning, nor more marking, just a different approach (p.4).

He reflects on how his experiences, initially as a result of involvement in the project and subsequent continued use of AfL, changed not only his pedagogy, but also that of teachers in his department. He also attributes the improvements in pupils' attainment to the resulting changes in classroom practice.

Over a five year period, following the introduction of AfL, Science exam results rose from below 40% A-C, to over 60%. Key Stage 3 results went up year on year, following a period of stagnation. Clearly learning improved over this period (Spenceley, 2009, p.3).

In this philosophy, the learner increasingly develops responsibility for their own learning, yet equally the teachers have a role in facilitating this development. Thus, if it is to be effective AfL is not seen as simply a set of tools and practices, which teachers can use in their lesson to help pupils more forward in their learning, rather it is seen as requiring a change in pedagogy specifically in relation to the roles and responsibilities between the learner and the teacher (Black et al., 2006; Spenceley, 2009; Marshall and Drummond, 2009).

This is particularly important in relation to the present research for the following reasons. Having made the decision to use the findings of Harlen (2004a) and Ofsted (2003b), as an instrument of analysis, to "identify" changes in teachers' assessment practice, it is possible to see, influenced by the SNS and Ofsted (2003b), the extent to which teachers adopted the tools for AfL into their classroom practice during the study period. However, as James (2006, p.2) cautions:

AfL practices can become mechanistic unless teachers understand the principles of learning on which they are based.

It is also worth considering, that even when teachers do understand and even embrace these underlying principles of learning, that the culture and climate in the schools, at both local and national level can act as barriers and inhibit the extent to which desired changes in teachers' practice can be achieved. Spenceley (2009) reports that despite this success with AfL for several years, there were a number of changes in his school overtime, which impacted on whole school adoption of AfL practice. These were at both local (change of head teacher and senior management team) and national level. He suggests that due to the ever-changing policy climate and performativity culture in schools, and the resulting range of initiatives introduced, teachers and senior managers:

> *... lost sight of the importance of AfL, and particularly of what AfL was originally all about. Emphasis moved away from formative classroom practice to a focus on learning objectives and three or four-part lesson plans. Lessons seemed to be judged more and more against a growing "tick list" of requirements, with AfL as just another box on the list (p.4).*

It could be argued that this perceived "tick list" approach inhibited teachers' adoption of the "spirit" of AfL, and steered them more towards adopting the "letter" of AfL approaches: tools to be used in the classroom and evidenced when being observed either internally by senior managers or externally by Ofsted. In common with the experience in PE, reported by Frapwell (2010), Spenceley (2009) also suggests that teachers' understanding of the original purpose of AfL became confused with tracking and leveling of pupils' attainment. He reports that in his school, as in many schools:

> *...individual student target grades began to over-shadow everything – often mistakenly thought of as 'being all about AfL'. Thus both staff and students began to concentrate more and more on the next level of attainment as a goal to be ticked*

> *off, losing sight of the methods by which reaching the target was to be achieved. 'Where next' took over from 'how next' (2009, p.4).*

With the benefit of hindsight, I now recognise that the limitations of my analysis framework in seeking only to identify where aspects of assessment practice are in evidence, for example, peer assessment, sharing learning objectives and self-assessment, do not allow for a more nuanced discussion in relation to teachers' commitment to the underpinning pedagogical principle of developing pupils' autonomy. Thus, whilst it is possible to evaluate teachers' changing practice, in terms of how they are adhering to the "letter" of AfL, it may be more problematic to assess if their lessons encapsulate the "spirit" of AfL (Marshall and Drummond, 2006, p. 133) and the extent to which their adoption of AfL practice reflects a "mechanistic" (James, 2006, p.2) approach.

However, at the time when the data was first collected, 2000, theoretical conceptualisation of AfL was underdeveloped. In their earliest work, on formative assessment (Black and Wiliam 1998a; 1998b) did not start from:

> *... any pre-defined theoretical base but instead drew together a wide range of research findings relevant to the notion of formative assessment. Work with teachers to explore the practical applications of lessons distilled therefrom (Black et al., 2002; 2003) led to a set of advisory practices that were presented on a pragmatic basis, with a nascent but only vaguely outlined underlying unity (Black and Wiliam, 2009, p. 1).*

It was not until 2006, coincidentally the final data collection point for my study, that Black and Wiliam developed an ad-hoc theorisation of AfL, which they have continued to develop in 2009.

Figure 2.1: Aspects of Formative Assessment

	Where the learner is right now	Where the learner is going	How to get there
Teacher	**1** Clarifying learning intentions and criteria for success	**2** Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	**3** Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	**4** Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	**5** Activating students as the owners of their own learning	

(Source: Black and Wiliam, 2009, p5).

Whilst Black and Wiliam (2009) would argue that all aspects of the model are essential, point 5 is particularly important if the changes in teachers' pedagogy, required by AfL are to be fully understood.

At this point, it is important to consider the context in which my study took place and its impact on PE teachers' assessment practice. In collating and publishing their notion of "Good assessment practice in PE", Ofsted (2003b) set an educational standard, against which the assessment practice of PE teachers would be inspected. Whilst it is easily possible to recognise the tools and methods associated with AfL within this definition, (Marshall and Drummond, 2006; Black et al., 2006; Mansell, James and ARG, 2009), the "spirit" of AfL is not so easy to interpret from this document. Thus, in such an inspection regime, where PE teachers were evaluated on the extent to which their practice adhered to this educational standard, it is to be expected that we will see a change in their practice towards this notion of 'good practice' as defined by Ofsted, regardless of whether they accepted or understood the underpinning principles of learning. Furthermore, taking into account the prevailing accountability and performativity culture in their schools, at the time of the study, we can expect to increasingly see these

changes at each of the data collection points, 2000, 2005 and 2006. However, whether this notion of 'good practice', promoted by Ofsted, is actually the most appropriate for dependable assessment in this subject, or whether the conceptualisation of PE, as encapsulated in NCPE (2000) is most appropriate for this subject, are beyond the scope of this research. The focus is to explore the changes in teachers' assessment practice within this context.

Summative assessment, or 'assessment of learning' is done periodically, at set times to summarise pupils' learning. This is usually at the end of a module, year or key stage in PE (Carroll, 1994; Gipps, 1990; Mawer, 1995). Whilst there is some variation in definition and practice, its commonalties include the following:

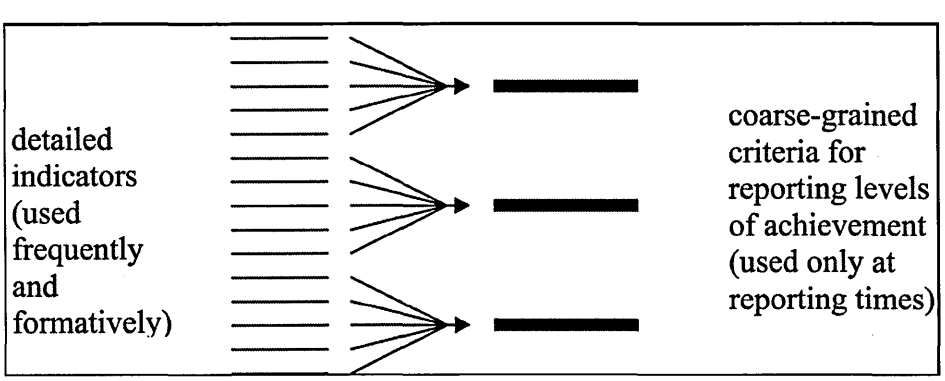
> *Summary assessment to establish the point a pupil has reached following a given period of teaching and learning:*
> *Specific assessment tasks, tests or exams administered outside the teaching and learning context:*
> *Summative performance of pupils' work assessed against specified criteria:*
> *Moderation of teachers' judgements, undertaken either on an internal or external basis (QCA, 2001, p.9).*

These definitions may lead to a misconception that 'assessment for learning' and 'assessment of learning' are two distinct processes, each with their own separate methodologies for assessment activities. However, their interrelationships are more complex. This was of interest to Harlen (2004a and 2006) who identifies that evidence, though collected through formative approaches to assessment could also be used to make summative judgements. She argues that the concept of summative assessment (assessment of learning) is a long established educational tradition, exemplified through testing approaches to assessment common in core subjects such as Maths and Science, whilst formative assessment (assessment for learning,) is a more recent development. She goes on to

suggest that using the terms 'assessment for and assessment of learning' can give the impression that these are two separate, discrete concepts, each with its own methods for gathering evidence. This is potentially one interpretation of the QCA summaries cited above. However, her central tenet is not that these are different types of assessment, rather that the essential difference is in the purpose of the assessment, or in her words 'how the information is used' (Harlen, 2005, p.105). However, whilst she argues that the essential distinction is the 'two conceptually different uses of evidence' (Harlen, 2005, p.106) she poses the question as to whether assessment evidence gathered for one purpose can also be used for the other? In this interpretation, the results of a test of learning could be used within an appropriate feedback framework to progress the pupils' learning. For example, they could revisit the questions they got wrong, analyse why and identify what they need to do to improve their answer in the future.

Given that formative assessment is by definition regular, frequent and ongoing, it could be argued that the information produced could be too detailed for meaningful use for summative purposes. Harlen (2005) suggested that indicators at two levels could be devised to address this issue, see Figure 2.2 below.

Figure 2.2. Formative and summative assessment using the same evidence but different criteria



(Source: Harlen, 2005).

This is important to the present research in that the levels of achievement, defined by Harlen (2005) as "coarse grained and used only at reporting

times", could equate to the NCPE (2000) levels. In contrast, either the detailed indicators could be developed from the learning outcomes of the individual lessons, or be a set of detailed assessment criteria developed from the NCPE (2000) levels by the teachers for sharing with their pupils. These might typify what knowledge skill or understanding is required to achieve each level.

Having reflected on the national reforms in assessment practice during the lifetime of the present research, this next section considers the role of teacher assessment in PE and reports on assessment requirements of the NCPE (2000).

Assessment in NCPE (2000)

NCPE (2000) has similarities with Sadler (1989) standards-referenced definition of the Australian approach to assessment, where standards-referenced assessment is typified by its use of general criteria, substantiated by descriptions of what pupils characteristically do at any given level. The programmes of study determine the knowledge, skills and understanding to be taught in any given key stage, while the level descriptions define the standards against which pupils' achievements are to be measured. These description are concerned with:

> *The types and range of performance that pupils working at that level should characteristically demonstrate (NCPE, 2000, p.42).*

At the end of Key Stage 3, teachers are statutorily required to make and record judgements against these standards and report them to parents. Whilst the need for teachers to arrive at an informed judgement of a pupil's knowledge, skills and understanding in PE at the end of Key Stage 3, necessitates a summative description at the end of year 9, the process of evidence collection may be undertaken throughout the Key Stage, in years 7, 8 and 9.

The obligation to gather such evidence neither excludes the use of either formative or summative modes of assessment, nor is there any assumption made that a formal rather than informal strategy must be adopted. Indeed all decisions regarding methods and approaches to assessment are left to the individual teacher's professional judgement, including decisions about the nature and purposes of assessment and the processes used to collect the required information, working within an individual school's assessment policy (SCAA, 1997).

Similarly, whilst teachers may want to keep their own records of pupil attainment, there is no statutory requirement to keep records on every pupil (SCAA, 1997). Thus, the responsibility for decision-making in assessment in PE is placed in the schools, with dependency on teacher assessment procedures. At the end of each Key Stage, teachers are required to reach a 'Best-fit judgement' of pupils overall attainment and progress:

> *Level descriptions are designed for End of Key Stage use only. Teachers will determine which level description 'best-fits' a pupil's performance (QCA, 1999, p5).*

This should be a summary of a pupil's whole profile of attainment in PE and may be based on evidence accumulated through out the key stage. If such summary judgements by teachers are to be 'dependable' (Harlen 2004a), then the extent to which validity and reliability that are considered during the on-going assessment process are central to their achievement.

Constructs of PE: Impact on assessment

In this section I will examine the impact of teacher's personal constructs of Physical Education on dependability in their assessment practice.

The dependability of teacher assessment relates to the knowledge, skills or understanding that one is seeking to assess. In this section, the extent to which PE teachers have a shared understanding of their subject and its

relationship with sport is explored, in order to consider how teachers' personal constructs of PE impact on the dependability of teacher assessment practice in this subject.

Shulman (1987) suggests that teaching begins with an understanding of what is to be learned and what is to be taught. This simple statement belies the complexity of the issues regarding curriculum choices. The NCPE (2000) provides a framework for a common PE curriculum in schools in England and Wales. However, whilst the NCPE (2000) sets out programmes of study, decisions about what is to be taught at each key stage are left to individual PE departments. In order to meet the requirements for the NCPE (2000) at Key Stage 3, within this framework schools are required to teach four learning strands. These strands are:

> *Acquiring and developing skills,*
> *Selecting and applying skills, tactics and compositional ideas,*
> *Evaluating and improving performance,*
> *Knowledge and understanding of fitness and health (NCPE 2000, p.23).*

For clarification, a copy of the NCPE (2000) Key Stage 3 Programme of Study is located in Appendix Five. These learning strands must be taught through four of the six areas of activity, which are gymnastics, dance, games, athletics, outdoor and adventurous activities and swimming. Decisions made in respect of the above choices reflect the relative values of both the PE departments and individual teachers in schools. Given the importance of teacher subject knowledge and its influence in what is taught, learnt and assessed, the range of choices available may result in quite diverse curricula across different schools. Thus, what is taught to individual pupils, depends on the subject knowledge and sporting experiences of their teachers or on the facilities available at their school.

PE is a complex concept particularly in its relationship to sport. Whilst it may be argued that the NCPE (2000) represents an agreed syllabus for PE,

there is a tension, as many commentators have observed, between those who argue the distinction between PE and school sport. Thus, the level of consensus is open to debate. This lack of consensus is well documented in the relevant PE literature (Bailey, 2005; Green, 2008; Lockwood, 2000; Murdoch, 1990; Whitehead, 2007). As far as it relates to the current research, a brief summary of the debate is provided. This is not however intended to be a comprehensive review of the literature in this field, more a clarification to inform the current research, the focus of which is on assessment.

There exists a general agreement that PE is about the development of physically educated pupils. However, there is significant disagreement about what this notion of being physically educated actually means. At one end of the spectrum, it is argued that the purpose of PE is to educate the pupils in terms of the knowledge and skills required to engage with the prevailing national and international culture of sport. This might be on many levels, from participant, in whatever capacity, through to informed observer (Alderson and Crutchley, 1990; Bailey, 2005). At the other end of the spectrum, the place of PE on the school curriculum is justified in terms of its capacity to educate pupils through the physical. Thus, for these advocates, PE is primarily valued as a process of learning, where the context is primarily physical (Murdoch, 1990; Whitehead and Murdoch, 2006; Whitehead, 2007).

As with all spectra, there are a myriad of views located in between these two extremes. Where the influence of the sport interpretation of PE model is most notable, is in the way examination level PE has developed at both GCSE and A level. However, the process model of PE in the main has dominated the development of core PE in this country for the last 40 years. Core PE is usually defined as non-examination PE. Currently, therefore, core PE is National Curriculum PE. DES/Welsh office reflects this process view of PE in its (1991, p.5) definition:

> *The purpose of this process is to develop specific knowledge, skills and understanding and to promote physical competence...the focus is on the child...rather than the activity.*

This view of PE, which underpins all previous versions of the NCPE (1992 and 1995), is explicit in the NCPE (2000). In terms of its significance for the current research, it is anticipated that an individual teacher's view, in terms of their level of agreement with the prevailing model, will have implications for the assessment process. This lack of consensus has implications not only for teacher assessment practice but also in terms of the credibility of the subject. As Whitehead (2000, p.7) articulates:

> *Working against the subject is the view that PE is recreation rather than education and therefore does not deserve its place in the schools curriculum. From this viewpoint, it is held that pupils have plenty of opportunities for recreation at break times and after school. Curriculum time should be used for serious study.*

Given this need to defend the subject and its place on the school curriculum, it is argued that a consensus about the nature and purpose of PE should be reached within the profession (Murdoch, 1990; Capel, 2000; AfPE, 2008). However, whilst this might be desirable in theory, in practice it has not yet been possible to achieve. This may be more easily understood when one considers the backgrounds of those attracted to PE as subject knowledge experts. PE as a concept exists only in schools and colleges. Outside these environments, those activities that feature as part of the PE curriculum are generally referred to as sports. It is therefore unremarkable that those who excel in this subject, and go on to train as teachers probably found more success in a sporting context than an educative one, for example achieving sporting honours at a regional or national level. These experiences influence their values and constructs of PE, which in turn may affect their assessment practice.

It is not my intention to critique the appropriateness of this conceptualisation; rather this debate is included to show the lack of consensus within the PE profession and its potential to impact on PE teachers' assessment practice. For the purposes of this study, it is accepted that the NCPE (2000) defines the construct of PE to be taught. Whilst it is clear that not everybody would agree with this conceptualisation, nevertheless, it is against the four learning strands of the programme of study (see Appendix Five) that pupils were to be assessed at the time of the research. This is significant for my work, as Ofsted focused their inspections of the assessment practice of the PE teachers, within this conceptualisation of PE. Thus, the Ofsted (2003b) notion of 'good practice' in assessment in PE, directly relates to this view of knowledge in this subject, and it is, therefore, within this policy context of Ofsted (2003b) and NCPE (2000) that changes in the teachers' assessment practice in Riverside Partnership are explored in Chapter Four.

Assessment in Physical Education: the role of teacher observation

Regardless of the purpose for which assessment is undertaken, it is necessary to ensure that assessments are dependable and accurately reflect a pupil's ability, at the time when it is undertaken. This requires the collection of appropriate evidence. Clearly, without evidence, judgements become a merely intuitive appraisal of pupils' learning, without any sound basis for the decisions made. Whilst it may be argued that PE teachers have always been engaged in formative assessment, where the primary purpose of assessment is in informing the teaching and learning cycle, the grading of pupils, in terms of their knowledge skills and understanding in PE is a relatively recent requirement.

The introduction of a National Curriculum for England and Wales in (1992) has raised the profile of assessment in PE. Influenced by the impact of GCSE and A level developments in PE and sport at that time, much attention was given by PE teachers to the devising of criterion-referenced

systems for assessing progress in NCPE, and quite complex procedures for recording the information were generated. However, even when the revised NCPE was introduced in 2000, little attention was given to the approaches for collecting the required assessment evidence.

Across the secondary curriculum, there is a variety of assessment methods available for use, by teachers, in the evidence collection process for both summative and formative purposes. These include tests, practical assessment, homework, projects, peer assessment and self-assessment. However, many PE teachers, indeed perhaps the majority, rely heavily on teacher observation of pupil performance to make summative judgements at the end of Key Stage 3 NCPE. Such teachers frequently argue that the strength of observation is its feasibility in the school context. The argument goes that assessment undertaken using this method is manageable in terms of time demands, in that it takes place in lesson time. However, as Harrison, Blakemore, Buck and Pellett (1996, p.42) point out:

> *If the assessment is to be thorough and truly useful, teachers should plan both data collection and procedures for recording information.*

In practice, most PE schemes of work are blocked on average over a 6-week half term. If one calculates the real time available in such a teaching context, once teaching and administrative tasks have been taken out, there is little time left over to undertake structured observations with an average class of 30 children. Clearly, such a strategy is limited by time in these circumstances.

A number of authors, exemplified by Carroll (1994), Mawer (1995) and Williams (1997) agree that observation is a useful method for collecting assessment evidence in PE as it can be effectively used within the teaching context without disrupting the structure of the lesson. However, the majority of such authors, unlike many teachers, see teacher observation as one component of an overall assessment strategy and not the strategy in total.

Much debate has taken place in the literature about the nature and purpose of authentic assessment (Gardner, 1992; Wiggins, 1989, 1998;Wiggins and McTighe, 2006). The argument promulgated is for ensuring that both teaching and assessment should be born out of real life settings. Authentic assessment can be defined as the assessment that is done in a real life setting as opposed to a more sterile testing environment. Thus assessing pupils' achievement in performing a forehand drive in Tennis would be better done in the context of the game, rather than in isolation. Wiggins (1989, p.45) identified four fundamental characteristics of authentic assessment:

> *Representative of performance in the field*
> *Criteria to be used in assessment should be taught to the learner*
> *Increased role of self-assessment in comparison to more conventional assessment approaches*
> *Students expected to present their work to demonstrate that their learning is "real" not perceived.*

The value of this approach is in its ongoing nature, embedded in the teaching and learning cycle and free from apparent interference by the assessment process. In this context, there are clearly arguments for the usefulness of teacher observation. There is some consensus that teacher observation can be a useful method in the collection of assessment evidence in PE. Indeed, many authors (Capel, Leask, and Turner, 2009; Mawer, 1995; Capel, and Piotrowski, 2000; Frapwell, 2010) support the view articulated by practising teachers, that assessment strategies, which rely on teacher observation, used in an informal or formal way, on an ongoing or summative basis can gather useful information, regarding the level and quality of pupil learning in PE. Such discussions focus on the teachers' professional skills, in terms of knowing their pupils' capabilities and their ability to use teacher observation to judge a pupil's progress. Thus, using observational strategies, a teacher is expected to assess what the pupils know, can do and understand in PE because of a given period of teaching and learning.

However, the dependability of relying on observation skills alone, receives little attention in the work of these authors. If its subjective nature is at all alluded to, then it is in the context of its positive contribution to the assessment activity, in that the teachers know their pupils well and the real life context increases the dependability of their assessment. Cohen, Morrison and Manion (1996), for example, argue that observation has several advantages in that it takes in the context of the situation as well as having a high level of validity and reliability. However, the problematic nature of such a strategy can be easily illustrated. For, whilst the conceptualisation of PE, represented in NCPE (2000), does include an emphasis on performance, the cognitive skills required for planning or evaluating may not be so easy to assess, using observation techniques alone.

However, whilst teacher observation clearly has some significant uses in the evidence collection process, the quality of the judgements made can be varied. Even though teachers have almost unlimited opportunities to observe student behaviour and attitudes, it does not necessarily mean their judgement will be objective and informed (Harrison et al., 1996).

Carroll (1994) supports this view, when he suggests that the quality of the teachers' abilities in the observational process is central to the validity and reliability of the subsequent summative judgements made. Whilst Carroll does raise issues regarding the quality of teachers' observational judgements, debate in this field is very limited. In order to find significant critique of the strengths and limitations of observational strategies, it is necessary to turn to methodological literature. Harris and Bell (1990) highlight the issue of subjectivity in observational method, suggesting that accounts may vary, even when several people observe the same event. This is a known concern in other contexts, for example the variability of witness accounts in legal cases.

If the process of observing is problematic, in terms of that which is seen by one observer may be different from that seen by another observer, so too are

the sampling methods implemented. Those who have discussed the sampling procedures for observational approaches to research (Denscombe, 1998; Wragg, 1994) tend to agree with the view put forward by Hammersley (1984). He suggests that whilst it is desirable to aim for 'intentional, systematic and theoretically guided sampling' (p.53) this is not always achievable. Compromises have to be reached and researchers frequently have to make do with ad-hoc opportunity samples. The argument then follows that the research could then be open to criticisms of bias.

This feature of the observational approaches to research has significance for the practice of teacher observation in the assessment strategy for PE. To explain, in the learning environment, the teacher frequently observes pupils' progress in PE on an ad-hoc opportunistic basis. In the same way that the researcher needs to ensure that the sample for observation is 'intentional, systematic and theoretically guided' (Hammersley, 1984, p.53) so too must the teacher ensure that observations of pupil performance are equally free from bias. This raises a number of questions. How can a teacher be sure that the performance observed is typical of the pupil concerned, a reflection of learning which has taken place as a result of their teaching or even an accurate demonstration of the pupil's knowledge, skill, or understanding in PE, as required by the National Curriculum?

Research methodological literature (Denscombe 1998; Yin, 2003) evidences much debate regarding the influence of the researcher as a person, in particular their values and beliefs and the impact of their presence on the subjects, when undertaking observations. These issues have implications in terms of subjectivity for the teacher observer in the context of pupil assessment in PE. For, just as the researcher is influenced in what to observe, when to observe, how to observe and why to observe, so too is the teacher. The subjectivity of such observation is encapsulated by Knudson and Morrison, (2002, p.96) who suggest:

> *The knowledge and expectations of the observer strongly influence what is observed.*

The subjective nature of teacher observation strategies is an important question for the current research. For even with so called objective criterion referencing, the decision regarding the extent to which such criteria have been met, still lies in the judgement of one person, the teacher. This view is supported by Wuest and Lombardo (1994, p.233). They found that:

> *The most common form of informal evaluation (teacher observation) is, at its very essence, a heavily subjective approach.*

'Best-fit' model for National Curriculum summative assessment

The subjective nature of assessment in PE is further highlighted when one recalls that the final summative judgement required by the NCPE (2000) is to be reached through a 'best-fit' approach. The implementation of the revised NCPE (2000), with its similarities to the Australian model of a standards-approach to assessment, (Macdonald and Brooker, 1997) has presented the PE teacher with yet another set of demands in terms of assessment practice. For, whilst standards-related assessment has similarities with criterion approaches, its validation of a 'best-fit' model is significantly different. Whilst it could be argued that the methodology of criterion referencing is predominantly objective, the 'best-fit' model is open to subjective judgements of individual teachers. To clarify, the levels of attainment are represented as developmental stages of pupil progression through the National Curriculum framework. As a result, teachers play an essential role in collecting evidence of pupils' achievements and interpreting this evidence in terms of the specified standards. Whilst levels of attainment were not implemented into PE until 2000, they were introduced in 1995 in other subjects. In 1996, Maxwell and Gipps reviewed teacher assessment practice in those National Curriculum subjects into which levels had been introduced. They reported that this represents:

> *... a substantial change from past educational practice, replacing the previous psychometric paradigm of assessment,*

emphasising measurement, scaling and formal standardised test, with the newer performance-standards paradigm, emphasising authentic, contextualised assessments and involving teacher judgement and interpretations of standards (p.19).

There is clear evidence of the influence of the reported psychometric paradigm on assessment practice in PE, which has grown out of the developments in GCSE and A level PE and A level Sport Studies. This contrasts with the introduction of a standards paradigm, through the formulation of levels in the NCPE (2000). This may lead to a shift in assessment practice in PE similar to that reported by these authors in other National Curriculum subjects.

The approaches to assessment used by teachers to reach summative 'best-fit' judgements are of interest to the current research. In the 1992 version of the National Curriculum, statements of attainment were used in all subjects to judge pupil progress and attainment. In 1995, in response to teachers' complaints that the sheer volume of statements of attainment made assessment against them unwieldy and unmanageable, level descriptions were introduced into some subjects. These were extended to all subjects, including PE, when the curriculum was revised in 2000.

Teachers are required to make 'best-fit' judgements against these level descriptions. The statutory advice to teachers for determining a level against the attainment target is to apply a 'best-fit' notion, which:

...is based on knowledge of how the pupil performs across a range of contexts, takes into accounts strengths and weaknesses of the pupils performance and is checked against adjacent level descriptions to ensure that the level awarded is the closest match to the child's performance in each attainment target(QCA / DfEE, 1998, p.8).

However, one problem with the attainment level descriptions is that they are broad, and lack specificity to be used as criteria for assessment. Gipps et al. (1998, p.6) observed:

> *If teachers are to use them for assessment purposes in anything more than a rough and intuitive way they may need to break them down: exemplars are also necessary in order to help classroom teachers make assessment against descriptions.*

In a paper presented to the American Educational Research Association (AERA) Conference in 1998, Gipps et al. presented the findings of two research projects, both funded by Schools Curriculum and Assessment Authority (SCAA) now known as Qualification and Curriculum Authority (QCA) into the role of teachers in National Curriculum assessment, undertaken in 1996 and 1997. Their findings, in relation to how teachers make 'best-fit' judgements of pupils' progress against the National Curriculum levels of attainment, are of particular interest to the present study. Their work, which included both primary and secondary teachers and headteachers, focused on the core subjects of Maths, Science and English.

Through the work for both studies, they found that the teachers used a variety of approaches to inform their 'best-fit' assessment judgements at the end of a Key Stage. These approaches are summarised in Table 2.4 below.

Table 2.4 How teachers make 'best-fit' judgements (1996)					
Number of Teachers	Y2 Teacher 60	Y6 Teacher 46	Head of English 34	Head of Maths 31	Head of Science 25
By making general 'best-fit' judgements	(43) 71.7%	(35) 76.1%	(18) 52.9%	(17) 54.8%	(18) 72%
By using 'best-fit' judgements in relation to children's portfolios	(35) 58.3%	(22) 47.8%	(23) 67.6%	(10) 32.3%	(5) 20%
By splitting the level descriptors (e.g. by creating separate statements and counting half or more as attaining a level)	(12) 20%	(8) 17.4%	(4) 11.8%	(5) 16%	(3) 12%
By identifying key aspects of a level description	(31) 51.7%	(23) 50%	(14) 41.2%	(8) 25.8%	(13) 52%

(Source: Gipps et al., 1998, p.7).

As can be seen from Table 2.4 above, responses varied, but there was a strong consensus amongst the secondary teachers of the value of averaging the set of levels, which pupils had attained at the end of the key stage whereas primary teachers reported using a general 'best-fit' judgement. In order to reach an understanding of how primary teachers defined 'best-fit' judgement, further work was undertaken with them in 1997. This is very relevant to the present research and their findings are set out in Table 2.5 below.

Table 2.5 How Y2 teachers interpret 'best-fit'	N =212
'Best-fit' interpreted as	Yes
The level description which overall describes the child's attainment better than the one above or below	(152) 71.7%
Must achieve 75% or more of the statements in the level description	(94) 43.3%
Must achieve important aspects of a level description	(55) 25.9%
Intuition	(36) 17%
Must achieve almost 100% or 100% of the statements n the level description	(32) 15.1%
Must achieve 50% or more of the statements n the level description	(4) 1.9%
Other	(3) 1.4%

(Source: Gipps, et al., 1998).

As can be seen from Table 2.5 above, most primary teachers appeared to interpret 'best-fit' in line with the QCA/DfEE guidelines of assigning the level that describes a pupil's attainment more precisely than the adjacent levels. However, Gipps et al. (1998) comment that this definition was added to the questionnaire at the specific request of SCAA, who were funding the research. Gipps et al. (1998, p.13) expressed their concern that it:

> *...does not tell us how the teacher makes the decisions as to what is 'appropriate'. In order to decide that one level is more appropriate than another, some judgement has to be made such as deciding key indicators or counting statements attained or alternatively intuition.*

Returning briefly to their 1996 study, Gipps and Clarke also concluded that whilst most teachers were quite sceptical about the 'best-fit' approach, they considered it an improvement on the system that it had replaced:

> *it was difficult to make decisions about pupils who appeared to fall between two levels and the notion of 'best-fit' was too vague. However having just been released from the previous system of counting the number of 'Statements of Attainment' a pupil had attained in order to determine a level, they said they found the approach more manageable so did not want it to be changed (p.12).*

This finding was of interest to the present study, where in contrast to the teachers in the studies reported by Gipps et al. (1998), the formulation of levels in the NCPE (2000) represented a transition from very wide-ranging statements:

> *working towards the expected level of attainment*
> *achieving the level of attainment*
> *working beyond the expected level of attainment (NCPE, 1995, p.20).*

to an 8 point numerical scale, with associated descriptions of attainments that were to be met in order for a pupil to achieve a particular level. Thus, in contrast to the teachers in the core subjects, this change was perceived by the PE teachers as a tightening up of the requirements for assessment.

In this literature review, I have examined a range of issues to contextualise the research questions posed by the present study. In Chapter Three, the research strategy is set out and the methodological decision-making that informed the study is reported.

Chapter Three: Methodology

This chapter sets out a rationale and justification for the use of a case study strategy locating the decisions made within the context of relevant methodological debates. It sets out what problems were faced, how these were addressed and the changes made in light of the developing research process. It details the ethical considerations for the study. It presents the lessons learned from the initial study and how they informed the main research. Finally, it reports on the methods used for data collection, how the data were analysed and the interpretive stance taken.

Why a case study?

Yin (2003, p.13) defines case study as:

> *...an empirical study that intends to investigate a contemporary phenomenon within its real life context, especially when the boundaries between the phenomena and the context are not clearly evident.*

The purpose of the present study was to gain an in-depth understanding of teacher assessment practice in PE and explore the changes to those assessment practices (2000 - 2006) within the Riverside Initial Teacher Training partnership. This study, as detailed in Chapter One, was undertaken at a particular time in a particular context. My intention was for "one aspect of a problem to be studied in some depth" (Bell 2005, p13). Therefore, case study was considered the most appropriate methodology for this research, primarily because of its scope, scale, context and time frame.

Yin (2003) suggests that the most appropriate strategy for addressing research questions involving the how and the why, as with my research, is case study. More so than the experiment or the survey, the case study offers opportunities, for comprehending the phenomena under investigation

holistically by combining information from several sources. Whereas experiments produce measured results from testing under controlled conditions and large-scale surveys using questionnaires or multiple interviews provide quantifiable data about the phenomena under examination, a case study seeks to achieve a different kind of understanding. An intrinsic case study was particularly suitable as the purpose of this investigation was to give a better understanding or explanation of a particular case (Stake, 1998), namely the changes in teacher assessment practice in the Riverside Partnership 2000 – 2006.

Further to Yin's definition (2003), it is argued that the value of a case study is to capture data for its uniqueness where the aim is not to infer findings from a sample to a wider population but rather to theory formulation (Bryman, 1988; Stake, 1998; Hammersley and Gomm, 2000). Teaching and learning episodes are unique to the teacher, the learner and the context in which each teaching and learning event takes place. These contexts shape teachers' practice, thus no two teaching and learning events, even if between the same teacher, the same learner and in the same school, will ever be identical. Rather than trying to control these conditions as 'background variables', Freebody (2003, p.81) suggests case study methodology acknowledges them as:

> *lived dimensions that are indigenous to each teaching-learning event. In that important respect, case studies show a strong sense of time and place; they represent a commitment to the overwhelming significance of localized experience.*

As my research sought to explore the changes in assessment practice of a particular group of PE teachers, who all work in initial teacher education in partnership with Riverside University, this focus on 'localized experience' was a significant factor in my decision to use case study methodology. It was important to my study that the teachers' assessment practice was researched in its own context, which is within their own school environment, in an attempt to ensure that an honest reflection of their

assessment practice was documented. The main variables in relation to the teachers involved in the study are

- Age
- Sex
- Number of years of teaching experience
- Institution at which their initial teacher training was undertaken
- Whether or not they were graduates of Riverside university
- Experience of continuing professional development activities in assessment in PE
- Role or responsibility within the Department
- Local Education Authority in which the teacher is employed
- Number of years experience at a particular school.

Controlling these variables in the present research would be both undesirable and impractical (Hammersley and Gomm, 2000).

The second reason for choosing case study methodology is that it dictates neither the paradigm, nor the methods the researcher must use (Hammersley, 1993; Denscombe, 1998). Much has been written on the issue of whether educational research should be quantitative or qualitative in character (Atkinson, Delamont and Hammersley, 1988; Denscombe, 1998; Winter, 2000). Whilst each paradigm has its own advocates, presenting them as precise and exclusive approaches, which in turn promotes an apparent conflict of paradigms, there are others who purport the usefulness of using quantitative and qualitative traditions in complementary, combined or mixed ways (Brannen, 1992; Reichardt and Rallis, 1994; Tashakkori and Teddlie, 1998). There are those who argue that it is oversimplified to talk about two clearly distinguishable paradigms. They contend that the differences within quantitative and qualitative approaches are no smaller or less significant than between them (see Hammersley, 1995). Finally, there are those who argue against the usefulness of the quantitative versus qualitative divide. They suggest that this approach to conceptualising research is unhelpful and does not reflect the reality of the research process

(see Ercikan and Roth, 2006). Whilst it was not possible to provide a detailed examination of all these debates, a brief summary of each has been included to contextualise the methodological decision making for the present study.

Early educational research demonstrated a strong commitment to the values of quantitative measurement, emulating the scientific approach to research. The ontological assumptions and philosophical basis of natural science are that the social world is external. It is interpreted as an objective reality, existing independently of an individual's conceptions, perceptions or experiences. This scientific method of enquiry draws from the school of positivism, specifically the verifiability principle of logical positivism that purports:

> *...something is meaningful if and only if it can be observed objectively by the human senses (Borg and Gall, 1989, p.17).*

However, critics of this positivist approach argue that it does not take into account the nature of human social life (Bird and Hammersley, 1995), the complexity of social interactions between individuals and the importance of understanding how people's perspectives shape their actions. The phenomenological philosophic premise, which underpinned the criticism of quantitative approaches to educational research and led to the development of qualitative methodologies, was one based on a subjective reality as opposed to an objective reality (Kvale, 1995). Thus, if there is a real world, it is different to everyone and can only be explained through the analysis of experience and interpersonal interactions. From this perspective, human behaviour is determined by phenomena of experience and may only be accurately studied through the development of idiographic methodology. Under this premise, the social world is not fixed but dynamic and changing, and this has to be taken into account in the research process (Marshall and Rossman, 2006).

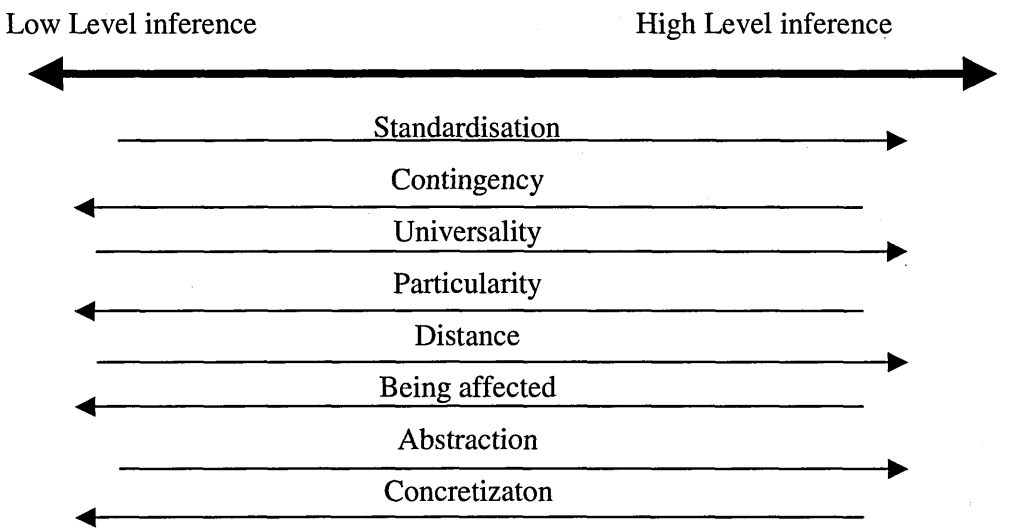
However, there is some consensus (Bryman, 1992; Brannen, 1992; Bird and Hammersley, 1995) that benefits can be drawn from both quantitative and qualitative methods. This suggests a need for a multidimensional approach to educational inquiry. Having analysed a number of papers offered as examples of good qualitative or quantitative research by proponents of a single approach, Datta (1994, p.67) concluded that, "the best examples of both paradigms seem actually to be mixed models".

Whilst this call for combining methodologies has received increased attention from researchers in recent years, it is not an entirely new debate. Patton (1980, p.20) implied such a view:

> *the debate and competition between paradigms is being replaced by a new paradigm – a paradigm of choices.*

Ercikan and Roth (2006) take this debate further. They argue that all phenomena and knowledge have, at the same time, both quantitative and qualitative facets. As a result, they reject the polarisation of research in terms of a quantitative and qualitative dichotomy and reject the related polarisation of concepts of objectivity and subjectivity. They also refute the notion that generalisability can only be applied to quantitative research. Instead, they propose a continuum, on which all research is placed with low-level inference at one end and high-level inference at the other with knowledge characteristics along eight dimensions. See Figure 3.1 below.

Figure 3.1. Continuum of low-level inference to high-level-inference research and associated tendencies for knowledge characteristics along eight dimensions.



(Source: Ercikan and Roth, 2006).

From this interpretation, instead of being different categories of research, quantitative and qualitative are located on different parts of the same scale and therefore only different by degree. Ercikan and Roth (2006) argue that this removing of the boundaries set by conceptualising qualitative and quantitative research as two distinct categories, allows the researcher to focus on the research questions as the main drive in determining the modes of inquiry rather than being limited by methods associated with a particular paradigm:

> *Instead of dichotomizing research into qualitative and quantitative, we need integrative approaches that provide the appropriate forms of knowledge (Ercikan and Roth, 2006, p.23).*

So, how did these debates influence my work? The origins of this study date back to 1998. Whilst my intention had always been to conduct a case study into assessment practice in PE, my original intent had been to adopt a positivist approach. At that time, I set out to prove the hypothesis that:

Teacher observation of pupil achievement is subjective and unreliable. Consequently, summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for PE based solely on an assessment strategy of teacher observation are invalid.

Reflecting perhaps, my own naivety as a researcher, I believed it would be possible to establish an objective, external reality, against which the assessment practice of a small number of teachers could be measured. However, as my study evolved, it became clear that such an approach was limited.

The principal decision to change from this positivist hypothesis approach to an interpretive investigation of a primary research question was a major shift for the present study. This revised approach related in part to a number of external factors that affected the study (detailed in Appendix One) and partly to the lessons learned through the initial study, detailed in later sections of this chapter. However, it primarily reflected my growing maturity as a researcher, including a better understanding of the value of "mixed models" (Datta, 1994) and recognition of the limitations of the scientific method for gaining insights into teachers' assessment practice.

Informed by these debates, in re-framing the study in September 2004, rather than adherence to a particular paradigm, finding the most appropriate way to address the individual research questions underpinned my methodological decision-making. As a result, I used both quantitative and qualitative approaches to data collection, analysis and interpretation, albeit not in equal measure. Case study is flexible and consequently capable of changing to take account of new insights or contexts. This was particularly important in the present study in accommodating the impact of this shift in focus, extended timescale and subsequent change to the methods used.

As detailed earlier, another reason for using case study related to the proposed scale and scope of the present research. As this style of enquiry is

particularly suited to the individual researcher, I considered the case study appropriate for a small-scale investigation such as this one (Denscombe, 1998). The flexibility afforded by case study meant that when I extended my data collection methods to include my students in the process, this change could be accommodated.

The research in this thesis is essentially exploratory, in that the purpose of my study is to explore changes in teachers' assessment practice in PE at Key Stage 3 between 2000 and 2006 within the context of the Riverside Partnership. My study did not seek to analyse pupils' assessment results in PE. Rather, it sought to examine the assessment process, what was looked for, why this was important and how this was being done, in order to explore how teachers' assessment practice changed in line with policy initiatives, which were the reality within which the teachers, teacher educators and student teachers were working at the time of the study, 2000 - 2006. In short, it sought to explore what was happening and if possible to offer reasons as to why this was so: therefore the main purpose of my study was to:

> *explore or investigate little understood phenomena or behaviours and discover the important underlying patterns, themes and factors which affect them (Falkner et al., 1999, p.17).*

Returning to my reasons for adopting a case study methodology, in addition to allowing the use of both qualitative and quantitative approaches to data collection and interpretation, a case study strategy also brings several other benefits to this research. Four such benefits are that case study research can:

> *offer rich insights*
> *allow multiple sources of information*
> *help identify further research needs*
> *identify new and fresh issues and insights for the research focus.*
> *(Simon, Sohal and Brown, 1996, p.32).*

In the context of the present study, it was my intention to seek "rich insights" into the research group, in order to gain a full understanding of their assessment practice at Key Stage 3, and changes therein between 2000 and 2006. "Multiple sources" of evidence were used to contribute to this understanding. "Further research" needs were identified and detailed in the recommendations section Chapter Four and "new issues" were added to the original focus of the investigation during the course of the study.

A final consideration in deciding to use case study relates to dissemination of the findings of my work. Not-with-standing any concerns about the generalisability of case study findings; a wide audience easily understands the results of a case study, they are immediately intelligible (Denscombe, 1998). This is particularly relevant to the present research, as the results are to be disseminated to teachers in schools, student teachers on initial teacher training courses and university-based academics.

The main criticism of a case study, in comparison to other forms of research, is that the results are not easy to generalise and whilst each case can "offer rich insights"(Simon, Sohal and Brown, 1996, p.32) into the particular situation studied, their application beyond the specific research setting is limited. Case study researchers counteract this in a number of ways. Marshall and Rossman (2006) suggest that generalisability corresponds to the positivist notion of external validity. They suggest that given the assumptions of case study analysis, where data is captured for its uniqueness, and where reality is viewed as subjective rather than objective, then the concept of generalisability in itself may be seen as problematic. Whilst there is some consensus that generalisability, is not a useful goal or standard for qualitative research (Stake, 1978; Goertz and Le Compte, 1984; Guba and Lincoln, 1985), the idea that the findings from case studies in one situation can be used to inform other situations is not universally discarded. Hammersley (1992) suggests that case studies are capable of producing general conclusions, which may be more widely applicable beyond the setting studied.

Schofield (1993) theorises that the applicability of case study findings depends on the descriptions of the case being studied and those situations to which generalisation is being made. Such descriptions are essential in allowing the search for similarities and differences between the situations. Denscombe (1998, pp.36 – 37) makes a similar point that:

> *The extent to which findings from the case study can be generalised to other examples in the class depends on how far the case study example is similar to others of its type.*

Analysis of these similarities and differences makes it possible to make reasoned judgements about the extent to which the findings from one study can be used to theorise about what might occur in another. This may be best summarised as the match between the study and other situations in which the concepts and conclusions could be applied and inferences made. This has similarities with the concept of transferability rather than generalisability considered by Guba and Lincoln (1989). From this perspective, case studies should be undertaken in such a way that to provide the reader with the information for them to decide if its findings are applicable to their own settings.

This notion of transferability was particularly helpful in my work in contextualising the wider application of my own case study findings. As a result, detailed descriptions of the case studied, the informants and the context within which the study took place have all been included in this report. By using instruments for data collection and analysis, which satisfy the criteria for being credible and dependable, and by providing the evidence and methods of analysis, which confirm that the study is free of bias, then my study can be seen to be trustworthy and reliable. In addition, by using multiple forms of data collection and more than one informant, the usefulness of the findings to inform other similar settings has been reinforced (Marshall and Rossman, 2006).

Before reporting on the design and conduct of the main research, I will now present a report of the initial study and how the lessons learned informed the main research.

Conduct of the initial study and how it informed the main research

The initial study, conducted between December 2000 and February 2001, was undertaken before the refocusing of the research in September 2004. As a result, much of it is no longer relevant and has been omitted from this thesis. However, there were a number of lessons learned through the approach and methodology of the initial study, which did influence the focus and design of the main research, and therefore the extent to which they informed the main study is reported on here.

At the time of its conception (1998), the purpose of the research was to test the hypothesis that:

> *Teacher observation of pupil achievement is subjective and unreliable. Consequently, summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for Physical Education based solely on an assessment strategy of teacher observation are invalid.*

In order to test this hypothesis a case study approach was adopted, which was to examine the practice of a small number of secondary PE specialists. The main purpose of the initial study was to further the methodology for the main research by identifying two appropriate subjects for this case study.

At that time, in order to achieve the requirements for quality in research, I felt that the case study must demonstrate both content and internal validity. For the purposes of the initial study, content validity was defined as the need for research instruments to sample adequately the domain they purport to investigate. Internal validity was defined as the extent to which alternative explanations for the observed effects may be excluded. The

extent to which the case study for the main research could meet these requirements depended, in part, on identifying appropriate subjects. These needed to be people who not only used teacher observation in the summative assessment process, for pragmatic reasons, but more importantly, on identifying respondents who were strong advocates for the validity of its use, and who genuinely did not see the need for other assessment strategies to be employed. Having established clear criteria for identifying such respondents, I designed a questionnaire, which I administered to all mentors in order to find suitable subjects for the case study.

The effects of 'researcher influence' are well documented (see Wragg, 1994; Cohen and Manion, 1980). This was recognised as potentially significant for the present research, even at this very early stage. Over a number of years, the group of teachers, from whom the participants were drawn had attended several mentor-training sessions specifically for PE initial teacher training led by me, where the focus had been on how to improve the PE initial teacher trainees' practice in assessment. Consequently, my very strong views about the over reliance on teacher observation in the summative assessment process in PE, and its implications for the preparation of valid end of Key Stage 3 reports to parents, were well known to the group. Given my role and prior relationships with the group, I was aware that this could result in the social desirability effect, whereby the participants consider agreement with the researcher's stance desirable. Honest response to the questions may be impaired if the respondents felt that to admit to sole dependency on teacher observation could be perceived by me as poor professional practice. Anonymous questionnaires could not be an option as the ultimate purpose of the questionnaire was to identify a sample of participants for future participation in the case study for the main research.

Having carefully considered this issue, I decided to use postal questionnaires, but which were adapted in the manner of their distribution. I distributed the questionnaires through routine moderation visits to the

partnership schools, undertaken in December 2000. My presence provided the teachers with the opportunity to clarify any misunderstandings on a one-to-one basis. However, they did not complete the questionnaires at the time of the visit, but they subsequently returned them by post. This afforded the teachers the opportunity to complete them in privacy later. Through this approach, I hoped to avoid problems of lack of clarity or researcher influence. Allowing the questionnaires to be returned by post resulted in a relatively low response rate (17 out of 26). However, I had anticipated this possible outcome and during a mentor-training meeting held in February 2001, I redistributed questionnaires to those mentors from whom no response had been received. Eventually, 25 out of a sample size of 26 were returned.

The design of the questionnaire, comprising of a series of closed and free answer questions, reflected my commitment to the merits of using a balance of quantitative and qualitative data to allow a more meaningful interpretation of the responses given. I incorporated the collection of qualitative data into the questionnaire to elucidate the quantitative data collected. Simple quantitative data was gathered through a small number of questions. A semantic differential rating response scale (Youngman, 1982) was adopted to permit easy comparison between answers. To avoid the risk of compartmentalising, open questions were then used to enable respondents to answer in ways that suited their frame of reference (May, 1997). The free response questions were included to give the respondents the opportunity to develop their own arguments in relation to each method. This in turn would generate more explanatory information.

The assessment methods examined were teacher observation, written test, peer assessment, pupil completed task cards and video recording. These were identified from a range of substantive PE literature, including Mawer (1995), Carroll (1994), Williams (1997), Capel (1997) and Capel and Piotrowski (2000), as being the most commonly used in the assessment of practical work in National Curriculum for PE at Key Stage 3. By placing equal emphasis on all methods, it was intended that any implied value

judgements of what is or is not considered 'good practice' would be avoided. (A blank copy of the questionnaire can be found in Appendix Six).

The initial study informed the main research in three significant ways. Firstly, the results challenged my perceptions about the dependency on teacher observation, as a sole assessment strategy in PE. For, despite success in identifying two suitable subjects, it was clear that overt dependency on teacher observation was not reported to the levels anticipated in the earlier stages of the research process.

This level of anticipation was warranted through my 10 years experience of working in initial teacher training in PE. Through the many forums of this work, including formal observation of teaching, informal discussion, training meetings and extensive student feedback of practice observed in schools, I had accumulated significant knowledge about assessment practice of a large number (100+) of PE teachers in secondary education. This led me to expect that a greater level of preference for teacher observation and a lower level of usage of other assessment methods would be identified, than that reported through the research process. The reasons for this discrepancy were of great interest to me. The use of teachers involved in teacher training may have influenced the results. It may be that the training they received as mentors in terms of developing trainees' skills in assessment may have influenced their own practice. Despite assurances of complete anonymity in the report writing process, the requirement for the teacher to give their name and school may also have influenced the nature of some of the responses given. However, it might also be that the changing culture in schools with respect to attainment and accountability, examined earlier in this thesis was already affecting teachers' practice. Whatever the reasons, I began to question the usefulness of the focus of the research.

The second key lesson learned from the initial study linked to my role in the research process or more importantly the perceptions of my informants of my role. The sensitive nature of the information required, coupled with the knowledge of the respondents of my strong views on what I considered to

be the questionable validity of a high dependence on teacher observation, may have impacted on the responses given despite the procedures detailed elsewhere to counteract this effect.

From this very early stage of the research, it became clear that my role in the initial teacher training partnership and that of my informants needed to be more fully considered, particularly in the potential ways that it could impact on the validity and reliability of the study. On the one hand, access to the schools in the case study was clearly facilitated, on the other a potential question linked to respondents' perceptions of being seen to promote 'poor' practice in assessment, either as perceived by me as Head of the partnership, by Ofsted or in the eyes of the trainee teachers with whom they worked was raised. As the study progressed, reflecting on my role in the study increasingly preoccupied me, and this is examined in more detail later in this chapter.

The use of a balanced approach to data collection in the questionnaire enabled me to both identify factual estimates of usage and preferences and gain an understanding of reasons given for these ratings. This affirmed my understanding of the value of using mixed method approaches to data collection and analysis. However, I realised that in order to attain the detailed understanding that I was seeking, I would need to place a greater emphasis on an interpretive approach to making sense of the data and I began to question the positivist hypothesis approach I had adopted.

Finally, although analysis of the results of the initial study demonstrated that two respondents met the criteria for suitability as subjects for the case study, once I revised the focus and methodology of the main research, they were no longer required.

Conduct of the main research

Informed by the work of the initial study, and a need to refocus the present research in 2004, I decided that the initial proposal to prove the hypothesis that:

> *Teacher observation of pupil achievement is subjective and unreliable. Consequently, summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for Physical Education based solely on an assessment strategy of teacher observation are invalid.*

was neither purposeful nor useful in gaining an insight into assessment practice in PE and that its value as a Doctoral level study was very limited.

In my continuing role in initial teacher education, I was acutely aware of the ways issues in assessment were being developed at national level, both in terms of practice in schools in general through the SNS, but also in PE in particular as reported on by Ofsted (2003). In consultation with my supervisor, I decided that a more meaningful study would be to move from an hypothesis approach, involving a very small number of teachers' practice, to an exploratory investigation of assessment practice involving teachers from across the Riverside Partnership. Whilst this fundamentally changed the nature and purpose of the research, I felt that the flexibility afforded by case study methodology could accommodate this change of focus and revised approach. Consequently, I devised the primary research question:

> *What assessment methods are used in Physical Education at Key Stage 3 in the Riverside Partnership and how have these developed between 2000 and 2006?*

Two supplementary research questions were devised:

In what ways do teachers of Physical Education in the Riverside Partnership consider the concepts of reliability and validity in their assessment practice at Key Stage 3?

How do teachers of Physical Education in the Riverside Partnership make 'best-fit' judgements as required by National Curriculum 2000 to decide on end of Key Stage 3 summative attainment levels, which are reported to parents?

Before going on to discuss the methods used in my study, and to detail the ethical considerations, it is necessary to define the boundaries of the case.

Defining the case

The case was defined as the Riverside Initial Teacher Training Partnership. The Riverside Partnership comprises Riverside University and 60 inner city, urban and rural schools, across the seven local education authorities, which surround the University. For the purposes of the present research, its membership is defined as the teachers in those secondary schools that were actively involved as mentors in the Riverside Partnership in the years 2000, 2005 and 2006. The number of schools used in each year depended on the number of student teachers in each cohort, which varied during the research period. The precise number of schools used in each year of the research is detailed on Table 3.1 below. However, defining the case as the Riverside Partnership in this way provided a framework, for this variable number of PE professionals to be part of the research process. Each respondent had a role within the partnership, that of school-based mentor for PE student teachers.

Data collection methods

In order to gain an in depth understanding of the teachers' assessment practice, I used a variety of data collection methods within this overarching case study strategy. These are summarised in Table 3.1 below.

Table 3.1 Data collection methods

Code	Method	Sample and size	Year	Approach
A	Questionnaires	Physical education Teachers across Riverside Partnership 25 Questionnaires	2000	Undertaken by the researcher for the initial study. Handed out during moderation visits and mentor training meeting.
B	Semi-structured interviews	Physical education Teachers across Riverside Partnership 18 schools, 40 interviews in total. 10 interviews 16 interviews 14 interviews Due to the scale and scope of the study, interviews from schools that only took part for one year were not included.	2000 2005 2006	Interview schedule formally devised by researcher. Interviews conducted by PE students in their placement school as part of their ITTE course.
C	Email Questionnaires	Physical education Teachers across Riverside Partnership 20 teachers	2006	Undertaken by the researcher
D	Interviews	Physical education teachers across 6 of the 7 LEAs across Riverside Partnership 6 teachers, 6 interviews in total	2006	Undertaken by the researcher

For the purposes of clarity, for the remainder of this report, each method is referred to by its coded letter, as detailed in Table 3.1 above. A brief

discussion of each method and its role in the present research is now presented.

A. Questionnaires

I revisited the data collected through the questionnaires as part of the initial study. Whilst the analysis had previously focused on identifying the subjects for the original case study, on re-examination, some of the data was useful in identifying the assessment methods in use across the Riverside Partnership in 2000 and proved valuable in allowing comparisons to be made to assessment practice in 2005 and 2006.

B. Semi-structured interviews between students and mentors

Before detailing the interview approach, the origins of this data collection method for my study need to be clarified. Since 1998, each cohort of students had been required to conduct a series of tasks, which examined assessment practice in their placement schools as part of their initial teacher training courses. These tasks were designed to increase students' knowledge and understanding of assessment issues in PE. They included a focused interview between student teacher and mentor and formal observation of two lessons to substantiate these interview discussions. Submitting written reports of each of these tasks was a mandatory requirement of one of the modules for the teacher-training course. This ensured that each student teacher completed the tasks.

Through these tasks, I had built up a large library of documentation about the specific assessment practice in each of the partnership schools since 1998. In re-focusing the study in September 2004, I recognised the potential value of such data for my own research. Its potential value was two fold. Firstly, the insights it could give into the practice of a larger number of schools in the partnership, and secondly, its role in addressing the concerns I had identified in the initial study about my role in the partnership and potential researcher influence. However, whilst one of the main strengths of using the student teachers in this way was that it allowed for the collection of a much wider range of data than would have been within the scope of a

sole researcher, the main difficulty lay with ensuring consistency in the data collection process. I will now detail how the interviews were conducted and the training given to each cohort of students to improve the quality of the data collection process.

The students were asked to conduct a semi-structured interview about teachers' assessment practice with their mentor. The format for the interviews was to represent as closely as possible a natural conversational approach. In order to ensure that they all covered the same areas, the students were required to ask the specific questions I had devised, (copy in Appendix Seven) and use this as the framework for the interview. However, respondents were to be allowed a degree of latitude within the framework, which would give them the "freedom to talk about the topic and give their views in their own time" (Bell, 2005, p.161). In order to achieve this, the students needed to use supplementary questions, as the conversational interview evolved.

In an attempt to keep the interview as naturalistic as possible, it was not necessary to record the interviews, as I felt the presence of the recording equipment could interfere with the natural flow of the conversation. Instead, the students had to make notes during the interview, which, on completion of the interview, they were required to type up under the headings provided by the questions I had set. The students were then required to share these notes with the interviewee, in order that the interviewee might check that they were an accurate record of the conversation. If necessary, the interviewee could make any amendments. Once agreement had been reached, the interviewees were asked to sign to confirm the accuracy of the record. The students submitted these accounts as an appendix to an assessment task, as detailed earlier in this chapter.

The first set of interview data used in this study was collected by the student teachers in 2000 as part of their university teacher-training course. At that time, its value in the research project had not been recognised and no additional guidance was offered to the students beyond their general

research methods module, which formed part of their initial teacher-training course. However, once the study was refocused in September 2004, its value in gaining an insight into teacher assessment practice in the Riverside Partnership was recognised.

In order to enhance the quality of the data collected, the students were given specific training in how to conduct semi-structured or focused interviews. This was done to try to enhance consistency in this aspect of the data collection process. Although this training still formed part of their research methods module, the content of the module was revised to focus on interview techniques. The students undertook a range of practice tasks to develop the necessary skills. These included practice interviews with their peers, using a set of questions, which I had devised and formulating supplementary questions as the practice interviews evolved. In this context, we examined how to avoid using leading questions and examined ways of ensuring accuracy in note taking and writing up the accounts. We spent a long time discussing how to reach an agreed account with the interviewee, particularly where perceptions of the discussions differed between interviewer and interviewee. In this way, these tasks mirrored the data collection method, which the students were to use in their schools to gather data for the present research.

The purpose of this training was to enhance consistency in this data collection method, the success of which relied on a large number of individuals employing similar practice. The training was repeated for all cohorts of students taking part in data collection for the study between 2004 and 2006.

These interviews (Method B) explored assessment practice in PE and in this context, how issues of validity and reliability, subjectivity and objectivity were addressed in assessment in the particular school in which the student teacher had undertaken a substantial teaching placement. This data collection method was particularly useful in giving the detailed 'rich

insights' (Simon et al., 1996) that this research was seeking, and proved particularly valuable in addressing research questions one and two.

However, even at the very earliest stage of the present research, I reported in the initial study that the possible implications of the effects of presence of the researcher on the respondents must be fully considered and appropriate strategies devised to take account of this. I was particularly concerned that my role as Subject Leader for the PE partnership and my known interest in assessment practice could influence the teachers' responses in the present study.

Having previously recognised the usefulness of the interview data collected by the student teachers in terms of their own professional development, I felt that the potential problems in terms of possible researcher influence might be addressed using this data collection method. To clarify, at the time the interviews were conducted, each student teacher had spent nine weeks in the school working very closely with their mentor and other members of the PE department. Over the period, the relationships between the student teacher and the departmental staff allowed the student teacher to conduct the interviews in a way that was perceived by the department as non-threatening. It also resulted in a higher degree of accuracy compared with an unknown researcher, in that the student teacher was a daily witness to the practice, which was the subject of the discussion. This proved to be a very useful source of information regarding both current and historical practice, and was analysed to investigate how assessment practice has developed between 2000 and 2005/2006.

C. Email Questionnaires

These were devised and sent to all mentors of placement two students. The specific purpose was to gather data regarding teachers' current practice in making 'best-fit' judgements in terms of National Curriculum for Physical Education levels at Key Stage 3, to address research question three.

As reported in Chapter Two of this thesis, in 1996 The Schools Curriculum and Assessment Authority (SCAA), now known as the Qualifications and Curriculum Authority (QCA), commissioned Gipps et al. to undertake a research project. They examined the consistency of teacher's assessment judgements in national assessment in England in the core subjects, Maths, Science and English in Primary and Secondary schools. The study, which they reported on in 1998, looked at how teachers were interpreting the requirement to make 'best-fit judgements' to allocate attainment levels for pupils at end of Key Stage 3.

My email questionnaire was designed based on the key questions identified by Gipps et al. (1998), to see if there were any similarities in the practice of PE teachers in the Riverside Partnership and the practice of teachers from the core subjects, who had taken part in this study. This data was also used to make further comparisons between the practice articulated in the semi-structured interviews B (mentors and students) and the interviews D undertaken by the researcher, in order to gain a deeper understanding of assessment practice in Riverside Partnership.

When responding to each question in the questionnaire, the teachers were asked to delete any of the statements that did not apply to them. These represented the elements that they never used as part of their assessment practice. They were then asked to rank the remaining elements, with 1 being those elements that most frequently featured in their practice. The lower the number allocated the greater the preference placed on the element. They could allocate equal marks to several elements to show equal usage.

D. Interviews with mentors

Semi-structured interviews were undertaken with mentors from a small number of schools. In choosing the schools for the interview sample, I considered two criteria. The first was that a student teacher had collected data from the specific school in 2006.

The second criterion was the Local Education Authority (LEA) in which each school was located. Riverside Partnership includes schools from seven LEAs. I had planned to interview a mentor from one school in each of the LEAs. As the purpose of the present research was to examine the practice across the whole partnership in assessment in PE, I felt that this would enable me to gain the widest range of information of all current assessment practice. However, due to staff illness one of the interviews was cancelled. Thus, six interviews were conducted.

The LEAs concerned range from rural to inner city. My purpose was not to examine any differences between LEAs; this would be beyond the feasibility and scope of the present study. However, I considered that the practice within each school might have been informed by any continuing professional development (CPD) in assessment offered to the staff in each school by its own LEA. Therefore, had I interviewed mentors in schools from only one LEA the information gathered could have been skewed.

I conducted these interviews after completing an initial analysis of the student interviews (Method B) in April 2006. As a result, they gave me the opportunity to probe further into the practice of each school in order to evaluate the extent to which assessment practice in each school in the sample reflected the Ofsted (2003b) principles of good assessment in PE and the recommendations for the use of formative teacher assessment for summative purposes (Harlen, 2004a).

Informed by the work of Nias (1989) I took a conversational approach and tried to make the interviews feel like open-ended discussions. I felt that this approach would help to minimise researcher influence, as detailed earlier in this chapter. The less formal the interview the more information it might yield. The semi-structured nature of the interviews allowed me to explore, though a series of open questions, aspects of each schools practice. Whilst I had identified a number of questions, I adapted the wording for each interviewee to allow me to ask the questions in the context of the normal flow of the discussion. The underpinning themes were those I identified as

relevant for the present study from the work of Harlen (2004a) and Ofsted (2003b). The interview schedule is included in Appendix Eight.

I chose not to use any recording equipment as I felt it would run counter to my efforts at achieving an informal natural discussion. However, I did make brief notes of key points during the course of each discussion. At the end of each interview, I then discussed my notes with each interviewee checking their agreement about the accuracy of what I had recorded, making any amendments as required. These were then typed up and a copy emailed to each interviewee. This gave them a final opportunity to clarify any details, as they felt appropriate. Five made no amendments with revisions received from one participant. In this way, my interview approach mirrored that required of my student teachers.

Before detailing the approaches to data analysis and interpretation used in this study, the ethical considerations for the study are now presented.

Ethical considerations

As Head of Riverside Partnership, the decision to investigate teacher assessment practice in this case was logical as access was unproblematic and freely given. Whilst all the schools were keen to host PE students on teaching placement, it was inappropriate to assume that the teachers would welcome close scrutiny of their practice in a research project such as this, as opposed to its scrutiny in the context of initial teacher training. Therefore, proper ethical consideration needed to be undertaken in negotiating access to undertake the study in these schools and in gaining informed consent from the participants.

In approaching the schools, and individual teachers, to be a part of the present study, I was initially informed by the (1992) statement of ethical guidance, issued by British Educational Research Association, (BERA) which was later revised in 2004 into "Ethical guidelines for educational research".

Firstly, the responsibility to the teachers, the respondents in the study, was a primary concern, particularly in securing their voluntary informed consent BERA (2004). At one of the regular mentor training meetings held at Riverside University in 2000, I explained the topic area and the nature and purpose of the research. Those mentors attending gave verbal agreement. A follow up letter detailing the research was sent to all mentors, and their formal permission was received.

From September 2004 onwards, when the focus of the study was revised, I sent an email to all mentors, each year informing them that the student tasks for their final assessment of their course were also one of the data collection instruments for the present research, Method B. I asked that any mentors, who did not want their data to be included in the research, should email me. In the three-year period, not one mentor emailed to withdraw from the study.

At every stage, when consent was sought, I assured all the participants that their data would be confidential and anonymous in the final report, and that they had the right to withdraw at any time. In writing up this thesis, the name of the university concerned and the names of all schools involved in the study have been changed to honour this promise of confidentiality.

As the student teachers were placed in the schools as part of an initial teacher-training course, formal agreements were already in place so no special action was required to negotiate access to the schools. As the study focused on teachers' practice, although the student teachers were required to observe lessons in which pupils were present, the permission for this aspect of the study was consistent with the permission required for them to observe lessons as part of their initial training course, therefore no extra consent was considered necessary by the schools and teachers concerned.

In order to conduct the interviews with a small sample of mentors (Method D), I negotiated access as part of a routine round of mentor moderation

visits. Again, mentors were fully briefed about the nature and purpose of the research. Issues of confidentiality were discussed, including their right to withdraw at any time, and consent was given.

In the final method used, that of the email questionnaires, Method C, mentors were briefed by email on the nature and purpose of the research, issues of confidentiality and their right to withdraw at any time. Following this briefing, those mentors who were happy to take part in the research returned completed questionnaires.

Data Analysis procedures

Having collected the data using a variety of methods, I needed to find the most appropriate ways for its' analysis. Given my commitment to the value of mixed methods, the choice initially seemed to be between content analysis and grounded theory approach. The differences between these approaches are summarised in Table 3.2 below.

Table 3.2 Differences: Content Analysis and Grounded Theory

Content Analysis	Grounded Theory
Bitty	Holistic
Go by frequency	Go by feel
Objective	Closer to the data, open much longer
Deductive	Inductive
Testing Hypotheses	Testing out themes, developing patterns

(Source: Easterby-Smith et al., 2002, p.345).

Having decided to move away from testing a hypothesis in September 2004, I decided that content analysis alone would not suffice. I had identified a number of research questions and, in applying the Ofsted (2003b) principles of 'good practice' and the key findings from Harlen (2004a), had established the themes I wanted to explore. This made the use of grounded theory alone equally less appropriate. However, some of the data generated,

for example from the interviews (Method B and D) did lend itself to a more intuitive approach where similarities and differences reported could be examined in more detail. Of the four sources of data, two (Method A and Method C) were most suited to simple numerical analysis, whereas one (Method D) was most suited to interpretive analysis. The data from Method B was analysed using a dual approach, which is detailed later in this chapter. Using this mixed approach to data analysis allowed me to gain a deeper understanding of teacher assessment practice in Riverside Partnership than would have been possible using either approach in isolation.

Before starting to analyse the data for my study, I devised a framework for analysis, which took into account, the key implications for practice identified by Harlen (2004a) and the Ofsted (2003b) findings and recommendations for good assessment practice in PE. Using this framework helped to give a strategic focus to the analysis of the data. Without it, the sheer volume of the available data was in danger of becoming unwieldy and overwhelming, which might have resulted in a more random reflection in relation to the research questions.

This framework was particularly useful in making sense of the data collected in the student mentor interviews (Method B). It consisted of eight headings derived from Ofsted (2003b) and Harlen (2004a). This enabled me to combine the "implications for practice," Harlen (2004a) and Ofsted (2003b) "good assessment practice in PE," resulting in a grid, against which to analyse each of the transcribed interviews (Method B).

1. Assessment Purposes
2. Assessment Types
3. Assessment Methods
4. Features of assessment practice
5. Conditions that affect dependability
6. On-going assessment
7. Involving pupils in the assessment process
8. Standardisation and Moderation

As the analysis progressed, I added a ninth heading 'views expressed' to address specific criteria that examined teachers' professional judgement and evidence of their consideration of reliability and objectivity.

Under each of the headings, I used sub-headings to identify reported evidence of what contributed to dependable or effective teacher assessment. A blank copy of this framework is located in Appendix Ten.

During my analysis, I also devised a scoring system that linked the evidence available from each of the interviews to the headings in the framework to indicate no evidence (0), some evidence (1), significant evidence (2) and evidence that this aspect was a significant factor in a teacher's or department's assessment practice (3). This was useful in interpreting the extent to which a particular assessment approach was being used in a particular school, for example was it embedded in practice or was it something that the department was considering implementing.

Having analysed the data from Method B against the framework, using this simple scoring system, I felt there would be merit in producing tables to give a simple comparison between assessment practice reported across Riverside Partnership at each of the data collection points, 2000, 2005 and 2006. However, due to the variable sample sizes in each year of the study, it was not possible to directly compare the practice reported using the raw scores I had assigned.

Having explored a number of mathematical approaches, I felt this would be possible using percentage agreement scores. I defined percentage agreement as the extent to which practice was reported across Riverside Partnership, where 100% is interpreted as evidence of embedded practice in all schools. The higher the percentage, the stronger the evidence in the data that this aspect of assessment practice was in use across Riverside Partnership. It should be noted that it does not refer to percentage of individual respondents using such practice.

The method for calculating each percentage agreement was as follows. Firstly, for each of the nine headings and their sub-categories in the framework, I needed to calculate the maximum score for each year. As 3 was the highest potential score: Maximum = 3 x N (where N = number of schools in each year).
For example in 2000, N=10. Therefore Maximum = 3 x 10 = 30

Secondly, for each of the 9 headings, and their subcategories, the total score given by all the respondents was then calculated for each year.
For example in 2000 Peer assessment score =11
The percentage agreement = 11/30 x 100/1 = 37%
Therefore, peer assessment in 2000 = 37% agreement.

I considered these scores to be a useful way of undertaking a comparison of the evidence for the assessment practice reported in each year of the study. Additionally, the compilation of the tables helped to make the approach to data analysis and interpretation more systematic. However, they were limited in reporting a detailed understanding of the changes in assessment practice. For example, the small numbers in each year meant that a change in only one teacher's practice could appear significant in terms of percentages (%) reported. This required a more inductive interpretation. Therefore, in order to gain a more meaningful understanding of the findings reported in the tables, vignettes from the detailed transcripts from Method B are also presented in the discussion and interpretation of the data in Chapter Four. These narrative accounts from the interviews conducted for Method B are used to inform the discussion about the nature and extent of the changes noted in teacher assessment practice and to examine reasons why particular practice was being reported.

The data from the email questionnaires (Method C) was analysed on two levels, using an Excel spreadsheet. Firstly, to show the incidence of commonality of each element used across the partnership. At this level, there is no indication of the order of preference of each element, simply the commonality of its use amongst the teachers in the Riverside Partnership.

Secondly, it was analysed to examine the order of preference of each element within the teachers' assessment practice. In order to make sense of the data, any element that had been identified as never used, (by being deleted before the questionnaire was returned), was scored at 1 mark more than the total number of elements in each question. Thus question 1 had 5 elements so deleted responses were scored 6; question 2 had 7 so a score of 8 was allocated and question 3 had 12 so a score of 13 was allocated. This enabled me to sort the responses in order of preference, whilst not skewing the results with elements that I knew were not part of a teacher's practice. I then compiled simple tables to show the outcomes of this analysis, which are then woven into the discussion of the findings in relation to research question three.

Finally, the data from the individual interviews (Method D) was analysed using an interpretive stance in relation to the framework devised. Extracts and vignettes from these have been inter woven through the discussion in relation to each research question. No mathematical manipulation of this data was undertaken

Before presenting the interpretation and discussion of the main findings of the study in Chapter Four, my role in the research process needs further reflection.

My role in the research process

I had been known to many of my informants for many years before my research began, and had freely expressed my opinions on assessment practice in PE and in my view its limitations, which was the very topic under investigation in the present research. Even at the outset of the present study, many of my informants were my former students. As such, they had attended my lectures and seminars on the topics of reliability and validity in assessment and the "problems" associated with assessment practice in PE. Others had attended mentor training events, again led by me, which focused on how to improve assessment practice in PE and in particular how to

improve the training in this area for each cohort of students on either the BSc. (Hons) Physical Education with Qualified Teacher Status or subsequently the Post Graduate Certificate of Education Physical Education (PGCEPE) courses.

As the seven years went by, an increasing number of my former students became mentors. Indeed, some of the students who had collected data in the earlier years of the study, 2000, 2001 and 2002 became mentors in schools in the Riverside Partnership, and were then themselves part of the sample from whom data was collected in later years, 2005, 2006.

I held a very powerful position in Riverside Partnership in that the decision about where to place student teachers year on year was entirely mine. The fact that many more schools wanted to host PE student teachers, than the number of trainees available, may have had an influence on some of the informants in the study. Put simply it was possible that there would be some respondents who would want to ensure that so called "good" practice in assessment was reported from their schools, otherwise their chances of receiving a student teacher in PE might be diminished.

In critically reflecting on this 'shared history', I needed to take into account where I stood in relation to my research informants, but even more importantly what my informants perceptions were in terms of this relationship (Hellawell, 2006). The concept of the "researcher influence"(Wragg, 1994), where it is recognised that some respondents will want to give the responses that they feel the researcher is seeking, the social desirability effect was initially useful to my reflections and had been considered carefully when setting up the initial study. However, as time progressed, and my maturity as a researcher developed, the uniqueness of this shared history, with its impact on assumptions and common knowledge, for example the assumption that teachers would 'of course be concerned not to rely on observation only, after all we all know that this is not considered good practice' required further conceptual deliberation.

Hellawell (2006, p.488) critically reflects on the concept of the "insider – outsider continuum" when considering the role of the doctoral researcher. The basis for his interest in this concept lay in his experiences of supervising a number of doctoral students. He was concerned with developing a degree of reflexivity on the part of the students on the impact of their role in the research process. A concern he expressed about the proposed work for one such student, which mirrors my own concerns about my impact on the present study, related to the power differential in a head teacher investigating the views and attitudes of his own staff, during a period of reorganisation at a time of financial constraint.

> *How could this head teacher interview his own staff and not simply receive a version of what the staff in question might surmise that he would want them to say, whether they genuinely believed it or not?* (p.484).

Through his paper, Hellawell (2006, p.489) reflects on what he calls "subtly varying shades of insiderism and outsiderism". This became increasingly important to me as my research progressed. Whilst the concept of insiderism and outsiderism in research can be dated back to the mid 20th century, see for example Gold (1958) who defines a spectrum from "complete observer" to "complete participant", Merton (1972, p.13) provides one of the earliest definitions of insider research. He concludes that within the sociology of knowledge there exists a:

> *balkanization of social science, with separate baronies kept exclusively in the hands of Insiders bearing their credentials in the shape of one or another ascribed status.*

Drawing on this work, Hellawell (2006, p.484) argues that the insider may be defined as an individual who possesses "a priori intimate knowledge of the community and its members". This definition was particularly helpful in reflecting on my own role in the present study.

Over the seven-year period of the research process, my "a priori intimate knowledge" of the schools in Riverside Partnership "the community" and the PE teachers and mentors within these schools "its members" increased year on year. Whilst by one definition I was an outsider to the partnership schools, on the other hand, I was very much an insider both in terms of my role as Head of Riverside Partnership and as the Course Leader of the Initial Teacher-Training courses from which many of the mentors had graduated. In summary, in my reflections on my role in the present research, I found strong agreement with Hellawell (2006, p.490) who suggested that:

> *There may be some elements of insiderness on some dimensions of your research and some elements of outsiderness on other dimensions.*

Understanding and reflecting on my "insiderness" informed both my methodological decision-making and my analysis and interpretation of the data collected for the present study.

In this chapter, I have detailed the methodology, methods and approaches to data interpretation and analysis undertaken for this research. In Chapter Four, the findings from all data sources are presented and discussed in relation to the research questions for the study. For ease of interpretation, some of the data is presented in table form, whilst other data is presented using a narrative style. In using this approach, I am able to both summarise and interpret changes in the PE teachers' assessment practice in Riverside Partnership between 2000 and 2005/6, within the policy context of the NCPE (2000) and Ofsted (2003b).

Chapter Four: Presentation and interpretation of data

This chapter presents the findings of the main research data. It reports on the interpretation of this data in relation to each of the research questions in turn and explores the findings of the present study in relation to issues raised in the literature review. For the purposes of clarity, I have structured this chapter around the research questions. Rather than presenting the data from each of the chosen methods in isolation, I am presenting my findings, interpretation and discussion from all methods used, in relation to each of the research questions in turn.

In this analysis, I shall comment on the similarities and differences, noted in the PE teacher's assessment practice: counting as interesting, both examples of any changes towards the Ofsted (2003b) notion of 'good practice' and the extent to which, the conditions that have been found to affect dependability in assessment in other subjects (Harlen, 2004a) are in evidence in PE.

Research Question One

> *What assessment methods are used in Physical Education at Key Stage 3 in the Riverside Partnership, and how have these developed between 2000 and 2005/2006?*

Having revisited the initial data (Method A) as detailed in Chapter Three, I was able to identify the range of preferred methods of assessment being used in PE, by the respondents at that time, and their reasons for these stated preferences. This provided an essential contextualisation of assessment practice, in the Riverside Partnership in 2000, with which any developments in 2005 and 2006 could be compared.

Table 4.1 identifies how each respondent rated each assessment method and shows each respondent's total usage of all methods. These have been placed in rank order, according to range of methods used, for ease of interpretation.

Table 4.1 Assessment methods used in 2000 for Key Stage 3 PE (Method A)						
Respondent	Teacher Observation	Written Test	Peer Assess	Task Card	Video	Totals *
I	5	2	5	3	5	20
B	5	3	4	4	3	19
C	5	2	4	4	4	19
R	5	3	4	3	4	19
A	5	2	4	3	4	18
Q	5	3	3	3	4	18
S	5	3	4	3	3	18
E	4	3	3	2	5	17
K	5	1	5	3	3	17
L	5	2	4	2	4	17
P	5	3	3	2	4	17
U	5	3	3	3	3	17
W	5	2	4	3	3	17
G	5	3	3	4	1	16
X	5	2	5	3	1	16
F	5	1	3	2	4	15
J	5	1	3	2	4	15
Y	5	2	2	2	3	14
D	5	3	2	2	1	13
H	5	3	2	1	2	13
V	5	1	3	1	3	13
N	5	3	2	1	1	12
O	4	2	3	1	2	12
T	5	1	2	2	2	12
M	5	1	2	1	2	11
**Totals	123	55	82	60	75	

(*Total score, by participant, for maximum rating of each assessment method used n =25).

**Max score for each individual assessment method used n = 125).

From the totals column in Table 4.1 above, it can be seen that a range of assessment methods was being used in 2000. However, teacher observation was the most highly rate method, scoring 123 out of a potential maximum of 125. It should be noted that whilst video assessment was also quite highly rated in the data collected through Method A, this too depends on a process of observation. The main reasons given by all respondents in 2000 for their level of usage of each of the assessment methods listed have been summarised in Table 4.2 below.

Table 4.2 Summaries of reasons given for higher levels (4-5) and lower (1-2) levels of usage for each assessment method.

Assessment Method	Summaries of reasons given for higher level of usage of each method	Summaries of reasons given for lower level of usage of each method
Teacher observation	Easy to implement in the limited time available Does not have a negative effect on performance as undertaken in a familiar environment for the pupils Valid and reliable Accurate Easy to assess large groups in limited time Moderation of teacher judgements User friendly Subjective method allows more frequent feedback Ongoing	None given as always identified as a high level of usage
Written test	Ensures that the understanding of all pupils is tested Summative exam to assess learning	Too time consuming Takes pupils away from practical activities Hard to access for less academic
Peer assessment	Provides information about two pupils at once, performance of the one assessed, and evaluation skills of assessor Gives immediate feed back to the performer more regularly than one teacher can for a class of 30 Addresses strands in NCPE (2000) Used to develop pupils' understanding Develops pupils evaluation skills	Time consuming to set up Unreliable, in that pupils may be over generous to their friends and over critical of those they dislike Lack of teacher experience Teacher more expert than pupils in assessment Demanding in terms of time spent training the pupils
Task cards	Provides a more objective set of data than pure observation Evidence collected can be held up for external scrutiny, such as parents and Ofsted	Lack of availability Too time consuming to prepare Slows down the lesson
Video recording	Available to refer to at later date Pupils can see themselves performing and improve their evaluative abilities Wet weather lessons Immediate feedback to pupils User friendly	Lack of equipment or technical support Takes too much time to set up Can have an effect on pupil performance as they are aware of the camera Lazy

From analysis of this preliminary data (Method A), it is clear that respondents in 2000 were aware of many of the issues concerning assessment raised in the literature review section of this report. For example, reference was made to validity and reliability. However, definition of these terms at this point was neither given nor sought. It is also interesting to note that the reasons given indicate a variety of purposes of assessment in evidence in the Riverside Partnership schools, with both formative and summative assessment being identified. Validity, reliability and the purposes of assessment are discussed in relation to research questions two and three later in this chapter.

Table 4.3 shows the rank order for each of the assessment methods reported:

Table 4.3 Rank order of assessment methods reported in 2000	
Assessment Method	Total Score (Max =125)
Teacher observation	123
Peer Assessment	82
Video	75
*Task Cards	60
*Written Tests	55

(* It should be noted that as the present research progressed, many schools across the partnership cited task cards as an example of written tests, rather than as a discrete method of assessment. As a result, these two methods were combined in the analysis of the data from Method B, and task cards as a distinct assessment method have not been reported.

Table 4.4 has been created to show the range of assessment methods, that PE teachers in the Riverside Partnership reported using in 2000, 2005 and 2006. Following the procedures, detailed in Chapter Three, the figures show the % agreement found through analyzing the data collected through Method B. This can be interpreted as the extent to which each method is being used within Riverside Partnership. It should be noted that the data from Method B suggests a lower level of usage of video assessment in 2000 than that initially indicated in the data collected in 2000 through Method A.

Table 4.4 Assessment methods	2000	2005	2006
Teacher observation	93%	83%	81%
Peer assessment	37%	54%	52%
Written assessment, e.g. task cards, tests	30%	27%	36%
Self reflection	17%	40%	48%
Question and answer	73%	71%	74%
Video assessment	7%	35%	38%

Whilst this table is useful in giving an overview of the changes in assessment practice that were noted at each data collection point, in order to fully explore these changes, each assessment method is now discussed in more detail.

Teacher Observation

This prevalence of teacher observation found in the data for the initial study, undertaken in 2000 (Method A) was consistent with the data collected through the semi-structured interviews between 2000 and 2006 (Method B). As can be seen in Table 4.4, there is a noticeably higher level of evidence of reliance on teacher observation, reported in each year of the study, in comparison to all other assessment methods reported. This is unsurprising and is consistent with the view that teacher observation is an important assessment method in Physical Education.

However, whilst it is clear that there is a preference for teacher observation, reported in each year of the study, this dependence is not so dominant in 2006 as it was in 2000. It is also significant that this is the only method of assessment that is reported as decreasing in evidence in teachers' practice; every other assessment method has increased through the whole period of the present research. This finding is interesting and raises a number of possible questions. Is the dominance of teacher observation reducing due to increased use of other assessment methods? To what extent has practice been shaped by developments in assessment at national level, with specific reference to the Key Stage 3 National Strategy? To what extent has the teachers' involvement with the Riverside Partnership influenced their practice?

There is evidence, in the teacher comments from the interviews (Method D) that seems to support the view that the changes at national level are making a difference to teachers' practice. This is particularly relevant to how teacher observation is used for assessment purposes. This is exemplified in the following comment from a PE teacher at Bellsunder School (2006):

Whilst I do use observation all the time in my teaching, I also draw on other methods, such as peer assessment and question and answer when assessing my pupils... My school recently ran a CPD session on using peer and self-assessment, which was very interesting, and I now try to include this where appropriate in my practice. This is particularly useful in Gymnastics and Dance sessions when they [the pupils] have been working on sequences... So yes, observation is important but using other methods to back it up makes me more confident in my judgements about kids' progress.

This response was not an isolated remark. For example, the following commentary from Goldvalley School (2000) shows how, in this school, the range of assessment methods has similarly developed between 2000 and 2006, with a dependence on teacher observation being prominent in 2000:

...the most frequently used method of assessment is teacher observation. It is seen as quick and effective method of scanning a large number of pupils as to whether pupils are successfully completing activities. Teachers can judge ability levels, pupil behaviour and task suitability through observation.

However, in 2006 a departmental commitment in this school to use a wider range of methods is reported:

Peer observation is included in PE lessons as often as possible. Non-participants are asked to observe their peers and give appropriate feedback. The department has a wide range of sheets that pupils can fill in if they are not participating which concern the assessment of a chosen pupil and their performance of a particular skill, set play or routine. For active pupils, they may be asked to observe their partner or group performing a practical task. Pupils always give positive feedback first and

then follow with constructive criticism in order for the pupils to gain an idea of what they need to improve for next time. This means all pupils can receive feedback, which rarely happens if the teacher is the only person assessing the class (Goldvalley, 2006).

This vignette is not an isolated account and is used here to exemplify the data presented in Table 4.4. This heightened awareness of the benefits of using a wider range of assessment methods was a common theme from the majority of schools within the partnership. The main reasons given for using a broader range of assessment strategies were linked to benefits of involving pupils in their own learning and were in line with the debates already presented earlier in Chapter Two of this thesis about formative assessment and assessment for learning. This would suggest that there is evidence that the work of Black and Wiliam (1998a) that underpins the Key Stage 3 National Strategy Assessment for Learning strand, is affecting teachers' practice in schools. This may be due to any or all of the following factors: continuing professional development, the focus on using a broader range of assessment methods by Ofsted inspectors when inspecting the schools or the PE department's involvement in the initial teacher training programme itself. This last factor, as previously discussed, must be fully recognised. Given the power of Ofsted, within the culture of performativity in the schools, it may be that the teachers want to appear to be using the 'best practice' Ofsted (2003b) which through their involvement as mentors in initial teacher training, they should be promoting to, and developing in, their PE ITTE students. To be seen to be doing otherwise might be interpreted as a weakness in their practice.

There was some evidence from the data collected in the interviews with students (Method B) that the teachers were reporting that a range of methods for assessment at Key Stage 3 PE was used. However, when teachers' practice was observed by the student teachers, they tended to see a greater reliance on teacher observation and question and answer than had

been suggested from the interview responses. The following vignette from Pineforest School (2005) is typical of schools where this was reported:

> *It is stated in the PE department policy document (2000) that the methods of assessment are:*
>
> - *Written materials (Record of Achievement)*
> - *Oral responses*
> - *Teacher observations*
> - *Student self assessments*
> - *Student peer assessments*
> - *Practical tests*
> - *Video recordings*

However, throughout a period of 6 weeks full-time teaching placement, where many teachers' lessons were observed, including two lessons specifically for the purposes of collecting data for the present study, the following comment from a member of the PE department at the school was reported:

> *A problem, however, with using these methods, is that they take time, need specialized equipment and specialized facilities. Teacher observations can be almost as effective and offer a more practical alternative (Pineforest, 2005).*

Returning briefly to the potential impact of being involved in the Riverside Partnership, it is also worth considering that this use of an increased range of assessment methods, reported in the 2006 data, may also have another explanation. Many of the mentors in 2006, were themselves trained through the Riverside Partnership and were previously my students.

Through my involvement in this study and through attending a number of continuing professional development activities about assessment in general, and in PE in particular, my own knowledge, and understanding of

assessment issues has developed. The prevailing performativity culture in schools also extends to teacher education. My involvement in leading Riverside University in three Ofsted inspections in PE, between 1998 and 2009 has also influenced my own perspective. Due to the direct link between Ofsted inspection outcomes and funding for teacher training places, I, like my schoolteacher counterparts, need to please Ofsted. As a result, my own practice has evolved in line with the Ofsted (2003b) notion. Whilst I might be able to offer a critical perspective to this construct, nonetheless the requirement to achieve success in Ofsted terms remains central to my professional career. Undoubtedly, both my academic development and my experience with Ofsted have shaped and informed the lectures, seminars and training events that I have delivered to both student teachers and their mentors in schools. This may have had some influence on the practice reported in the Riverside Partnership, either on those mentors who attended the meetings or on the mentors in 2006, who were themselves students in the earlier years of the present study from 2000. However, whilst it might be rewarding to believe that in some small way, my work has influenced teachers' practice, there is clearly not sufficient evidence to substantiate such a claim. Nevertheless, a factor that should not be ignored in seeking to interpret the data gathered for the present research.

There is evidence from the present research to suggest some connection between using a wider variety of assessment methods and a decreased justification of a teacher's professional judgement as the sole methodology for assessing a pupil's attainment.

Table 4.5 Views expressed Professional Judgement	2000	2005	2006
Justification of reliance on teachers' professional judgement	70%	19%	29%

In considering Tables 4.4 and 4.5 together, it can be seen that as teachers' reports of using a wider variety of assessment methods increased (see Table 4.4) the frequency of teachers reporting that they relied solely on their 'professional judgement' to determine a pupil's attainment level decreased,

(see Table 4.5). This may indicate an increased skilfulness in assessment practice.

In 2000, it was frequently reported that whilst it could be argued that teacher observation was subjective, this was counteracted by the professional experience of the teacher conducting the assessment. The following commentary from Goldvalley School serves to illustrate this point:

> *The response from this discussion, regarding the issue of teacher observation being open to subjectivity was focused on the professionalism of the teacher that is assessing and observing the class [...]. The teacher did not deny that the methods of assessment that were used in the PE department could be regarded as unreliable and subsequently lack validity [...] but then again what method is regarded as totally foolproof. The teachers were all fully aware that assessment is an area for development within the department and that new practices and processes were due to be employed soon (Goldvalley, 2000).*

This comment seems to contradict some of the reasons given for using teacher observation in the data collected, (Method A) for the initial study in 2000, where it was argued by the majority of respondents that teacher observation was both valid and reliable, as it does not interfere in the teaching and learning process. This contradiction is one of many found in the data, which, given the number of schools in the Riverside Partnership, reflects the diversity of assessment practice and interpretation of assessment issues. This comment could also suggest that as Ofsted (1998 onwards) was highlighting the problematic nature of assessment practice in PE, some teachers were starting to recognise these concerns in their own practice, even if they had yet to find solutions. It is important to note that the value of teacher observation was not being rejected either by Ofsted. Nonetheless, raising awareness of its limitations and discouraging reliance solely on a strategy of teacher observation to assess in PE was high on their agenda.

In the data collected for the present study, there was also some association between the recognition of the need to consider the validity and reliability of the methods used and the justification of professional judgment.

Table 4.6 Views expressed: Validity and Reliability	2000	2005	2006
Justification of reliance on teachers' professional judgement	70%	19%	29%
Validity considered	55%	84%	86%
Reliability considered	45%	81%	86%

The data suggests that, as recognition of the former increased, dependence on the latter decreased (see Table 4.6). This is further explored later in this chapter in relation to research question 2.

The reasons for this change in practice, in relation to dependence on teacher observation, are complex to determine. However, this change between 2000 and 2006 is interesting in that it is reported so widely among the participants, across the Riverside Partnership.

Written assessment

The use of written assessment was reported in the initial data (Method A) as the least commonly used form of assessment for PE at Key Stage 3, closely followed by Task cards (Table 4.3). As already stated, as the present study progressed, Task cards were commonly cited as examples of written assessment, rather than as a distinct assessment method. Therefore, they are not reported separately in the data collected through Method B. From Table 4.4, it can be seen that there was a slight decline in the use of written tests from 30% agreement in 2000 to 27% agreement in 2005. By 2006, there was some increase in its usage to 36%. However, the changes in these figures are so small they are barely significant. This finding may be interpreted in a number of ways. As discussed in Chapter Two of this thesis, assessment in PE at Key Stage 3 rose out of the more formal modes of assessment that had been developed for GCSE and A level PE relatively recently in the late 1980s (Green, 2008). Much of this debate is beyond the

scope of the present study. However, this formal approach did initially influence the use of written tests in PE assessment at Key Stage 3. This was compounded by the perception that having some written evidence of pupils' learning in any context, regardless of whether it was the most appropriate way of assessing that aspect of a pupil's learning, would be considered by Ofsted as good assessment practice, when examined in an inspection. This is substantiated in the reasons given for using written assessment in the initial data (Method A) in relation to Task cards, the most commonly used form of written assessment in PE:

> *Evidence collected can be held up for external scrutiny, such as parents and Ofsted.*

The following commentary from Churchenfield School is typical of the use of written assessment tasks that were reported in PE in 2000:

> *There are times when the pupils assess each other and provide feedback – both verbally and written forms, e.g. in gymnastics, where they have to complete evaluation sheets about the performances of other groups.*

It could be argued that the process of evaluation was the key factor in improving pupils' learning, while the resulting written product served other purposes.

The following reasons were articulated in the 2000 data (Method A) against using written assessment:

> *Too time consuming*
> *Takes pupils away from practical activities*
> *Hard to access for less academic.*

These could provide the explanation for its limited use within the schools in the Riverside Partnership, in assessment at Key Stage 3 Physical Education.

Evidence from the interviews (Method D) substantiates this view. For example, the PE teacher at Churcholt School commented:

> *I use written assessments very rarely now [...] In my work in SEN department, I realised that often the less able pupils know the answers the teachers are looking for but find it difficult to write them down [...] This has made a difference to my work in PE [...]. I might give them a task card, but I make sure that it has lots of pictures or diagrams to help the pupils understand the task, and I try to use tick boxes where possible so that writing is kept to a minimum. Problem is, it takes twice as long to do differentiated task cards, so although I know I should do them, I don't really have time (Teacher, Churcholt School, 2006).*

However, in the other interviews (Method D), where written assessment was mentioned, the teachers usually referred to pupil assessment profiles or portfolios of attainment. This was also found to be the case in the data collected through Method B. Therefore, whilst profiles and portfolios of attainment are not examples of the types of written assessments this study set out to research in 2000, these are now discussed in relation to the data collected.

This use of profiles and portfolios of attainment was noted in a significant number of teachers' practice across the Riverside Partnership, in a variety of formats. The following commentary from Hurley School (2005) exemplifies one of the many formats reported:

> *The PE department uses an assessment booklet to record and assess pupil's attainment and progress in all areas of the National Curriculum (Hurley, 2005).*

There was evidence in the data collected that some schools are using these profiles to serve a number of purposes. For example, in Rivermeadow

School, it was noted that such 'profiles' not only record progress but also engage pupils in reflecting on progress as well as reporting progress at appropriate times through out the year:

> *Students are encouraged to reflect on, and record, their own progress through the 'student profile' sheet, which is a self-assessment of each activity throughout the year. These are kept in the form file, which is available in the Sports College Office. Students also record their overall progress on the school subject report (Rivermeadow, 2005).*

Issues raised in relation to assessment practice in PE by Ofsted (2003b) could have influenced the development of such profiles or portfolios. Ofsted (2003b, p.6) suggest that:

...the most effective assessment is linked to the National Curriculum programme of study; precise learning objectives are described in language that pupils understand. Teachers have an agreed view on what constitutes performance at every level.

From the data collected, there is evidence that in some schools across the partnership, assessment and attainment profiles or booklets in PE are being devised to help teachers meet these recommendations for good assessment practice in PE. Thus whilst the evidence is mixed in quality, and does not come from all schools in the Riverside Partnership, it should be noted that some are including in such profiles or booklets, agreed criteria written in pupil friendly language. From the evidence collected (Method B), it seems that at Key Stage 3, these criteria most commonly relate to the end of Key Stage attainment levels, with limited evidence in the data, of the inclusion of learning outcomes by a few schools across the Riverside Partnership. The role of shared criteria is more fully discussed later in this chapter.

Linked to this interpretation of written assessment, which was noted in the data as the study progressed, is the balance of ownership between teachers

and pupils of written records of attainment and progress. Central to this is what I have termed a process versus product model. To explain, some schools have developed what might be termed a process model, in that by engaging the pupils in reflecting on and recording their own attainment against shared criteria that pupils understand, they are involving them in the process of assessment. In contrast, others have developed what could be termed a product approach, where the teacher records the results of any assessments undertaken, for example the levels or scores awarded.

From the data collected for the present study, there is evidence to suggest that even in schools where the staff-owned database (product) approach is adopted, pupils are still being engaged in their own assessment, in order to attain these levels that are being recorded by the teacher. However, the extent to which the teachers regarded the recording process itself, as a further opportunity to engage their pupils in self-reflection, (process) or regarded it as an administrative duty linked to record keeping and accountability (product) is of interest.

Data from the present study would suggest that the majority of schools in the Riverside Partnership have adopted a product model. An example of this can be seen in the commentary from Croft School (2005). This vignette serves to illustrate how teachers can interpret formative assessment solely from the perspective of teachers' practice, without ever really involving pupils in either the process of assessing or recording their attainments:

> *...formative assessment is carried out by the teachers' constantly, this helps teachers to build up profiles for pupils and aid in assigning key stage levels at the end of the unit of work. The pupils were not informed about the assessment, as the pupils are aware that assessments happen every lesson. The teachers then record their assessments on the departmental database (Croft School, 2005).*

The use of the word 'assigning' in relation to Key Stage 3 levels is particularly interesting, in this context, as it suggests that this assessment is 'done to' the pupils rather than their 'being engaged' in the process. This raises an interesting question about the balance of responsibility for learning, between teachers and pupils, as discussed in Chapter Two.

Marshall and Drummond (2006, p.133) suggested that the "spirit" of AfL, was central to effective formative assessment. From this perspective, the teacher's role in developing pupil autonomy is the underlying pedagogic principle for successful AfL. They suggest that in classrooms where this is in evidence, there is an increasing shift in responsibility for learning from the teacher to the pupils, as they become more autonomous. However, in interpreting this vignette from Croft school, it appears that whilst the teachers claim to be engaging in formative assessment, the lack of pupil involvement in the assessment practice indicates that the teachers have retained full responsibility for the learning and assessment process. Rather than enabling their pupils to take more responsibility for their own learning, there is a total removal of responsibility from the pupils to the teachers, in that they are not even informed of assessments taking place, nor are they included in deciding on the assessment outcomes.

The comment that, "...the pupils are aware that assessments happen every lesson" (Croft School, 2005), indicates that informal assessments are being undertaken. However, it may be that informal assessment is being confused with formative assessment.

At this point, it is worth reflecting on the practice that existed in 2000 and commenting on how it has changed in relation to this 'process versus product' model over the period of the present study.

Table 4.7 Pupil recorded attainment and staff recorded attainment	2000	2005	2006
Progress and attainment recorded in pupil owned progress file, booklet or record of achievement	30%	31%	46%
Progress and attainment recorded in staff controlled database	10%	38%	64%

The most significant increase is reported in relation to teacher owned methodologies, (product) from 10% agreement reported in 2000 to 64% agreement reported in 2006. In comparison, pupil owned methodologies, (process) stayed constant between 2000 and 2005, with a small increase to 46% in 2006. There was some limited evidence of schools developing both methodologies in parallel. For example, County Springs School (2005) reported using "...a departmental database and pupil profiles". However, overall, this data would suggest that during the period of the present research, the product rather than process approach, received greater attention from the PE teachers, in the Riverside Partnership schools.

In further interrogating the data, it reveals that in 2006 all schools in the sample, except one, reported using some form of teacher owned database to record attainment, in comparison to only one school in 2000. This change in practice may simply be explained, by the wider developments in information technology. Between 2000 and 2006, there were a number of government initiatives to promote the use of computers in schools, not specifically in PE, that occurred in education, for example "Curriculum Online" or "Computers for Teachers". The first was an initiative to provide free access to a range of software for schools; the second was an initiative to provide computers for teachers at significantly subsidised rates. However, given that there was variation in the extent to which the Riverside Partnership schools reported using such databases, it is interesting to explore further. Only two schools reported that their use was an essential part of the assessment practice of the department and both schools reported receiving

positive comments about their systems for recording by Ofsted, during inspection visits. The comments reported, linked specifically to tracking pupils' progress. One school, Wetland, explained that they were developing their assessment system in partnership with their LEA and a commercial company. In their advertising material, they claimed it was "endorsed by Ofsted". It has been available, as a product, for other schools to buy, since 2006. During my interview with the PE mentor from Wetland School, when this system was discussed, I asked how the grades, that were recorded in the system by teachers, were decided and what role the pupils played in the grading process. She reported:

> *The teacher who teaches the activity assigns the grades at the end of the unit of work [...] the pupils do some formative assessment in the lessons, but the grades are awarded by the teacher (PE mentor, Wetland School, 2006).*

Whilst this example is only from one school in Riverside Partnership, it does raise some interesting questions. Given that the grading system is maintained entirely by the teachers, does this separation between the methodologies for collating attainment and the assessment process affect the learning of the pupils? Are opportunities to use AfL being missed, in favour of concern about numerical manipulation of scores? Did the development of such systems, contribute to the what Frapwell (2010, p.13) sees as:

> *...the [PE] profession's obsession to convert every bit of progress a learner makes into a number [level] or a grade to create data.*

The use of these recording systems is further examined in relation to research questions 2 and 3 later in this chapter .

Video assessment

The reported use of video assessment has increased throughout the present research (see Table 4.4) with the most significant increase noted between 2000 and 2005, from 7% agreement in 2000 to 35% agreement in 2005. There was then a slight increase noted between 2005 and 2006, from 35% agreement to 38% agreement. Whilst this increased use of video is unsurprising, as it mirrors the growth in technology in society as a whole, the fact that it has not increased further is worthy of discussion. Given the very practical nature of PE, and that each performance is unique and fleeting, video evidence could be used in a variety of ways to engage pupils in their learning and assessment. In 2000, the reasons given for not using video assessment included:

> *Lack of equipment or technical support*
> *Takes too much time to set up*
> *Can have an effect on pupil performance, as they are aware of the camera*
> *Lazy teacher*

Despite the fact that during this period, on a wider societal scale, technology was becoming cheaper and more widely available, similar concerns were still in evidence in the data collected in 2005:

> *A problem, however, with using these methods are that they take time, need specialist equipment, staff training and specialist facilities. They also go wrong then the kids get bored and mess about. Teacher observations can be almost as effective and offer a more practical alternative (Pineforest School, 2005).*

It may be that the last sentence, in this commentary, is more illuminating of the reasons why the use of video assessment has not further increased, throughout the Riverside Partnership schools. This may also indicate a lack of confidence, or technical ability, on the part of the teacher, in the use of

video assessment, particularly in comparison to their pupils. To clarify, the current generation of pupils are the most technologically literate, with most having access to digital cameras and videos, often as a feature of their personal mobile phones. Consequently, most are used to both using technology and seeing themselves recorded on camera, albeit that there may be individuals who are still not at ease with the experience. Thus, despite the cost of technology having reduced significantly over the whole period of the present study, whilst some increased use of video assessments is noted, the data from the present research suggests that it is not used widely across the Riverside Partnership.

Returning to the schools that do use video assessment, there is evidence that the function of such assessments is varied. One purpose reported is that of providing immediate feedback to the learners. This can then be used to inform their learning, which is consistent with Ofsted (2003b) notion of good assessment practice in PE:

> *...using ongoing assessment evidence to provide specific feedback to guide pupils towards improvement (p.1).*

An example of the use of video assessment for feedback purposes, is cited below:

> *video analysis [...] provides a good source of evidence and feedback. The pupils learn more when they are shown their performance, which allows them to evaluate each other and moderate with the teacher (Pineforest School, 2005).*

This link to the potential role of video assessment to both self-assessment and peer assessment is discussed later in this chapter .

Another use of video assessment, reported in the data, was that of providing the teacher and the learner with a recorded performance upon which to base their assessment judgements. This can then be viewed multiple times, as

required, both during the lesson by pupils and after the lesson by teachers. This attempts to address the fleeting nature of performance in PE. The following extract from the interview (Method D) with the mentor at County Springs School (2006) exemplifies this practice:

> *During the year 8 boys gymnastics lessons, I sometimes use trampettes. In one lesson, I set up two lots of apparatus, [trampettes and mats], one at each end of the gym. At one of them, I set up a video camera, linked to a laptop that recorded their performance. The boys could watch themselves, to see how well they are doing and how they could improve. They were also able to rewind the recordings and compare each of their attempts [...] and I was able to look at the recordings after the lesson. It worked really well but there were only about 11 of them in the class (Teacher, County Springs School, 2006).*

The teacher did go on to say that, whilst this was a very effective lesson due to the small class size, it would probably not work so well using video assessment in this way, if the class size was significantly increased. However, he did say that he has since used these recordings in lessons with larger groups to show examples of very good performances as well as typical errors in the skills being learned.

The final use of video assessment reported, was that of providing evidence, which could then be used by teachers, within the PE department, in their moderation and standardisation procedures:

> *... a number of teachers used video analysis to evaluate pupils and standardize the grades given for a specific level of performance (Mansion School, 2005).*

The reasons given for using video assessment in this way were linked to issues of reliability and validity. These are explored in relation to research question two, later in this chapter .

When considered in light of the issues raised in this discussion, whilst the increased use in video assessment is unsurprising, its relatively low level agreement of 38% in 2006 is unexpected.

Question and answer

From the data collected for the present study, there is evidence that the use of question and answer, for assessing learning has remained constant throughout the schools in the Riverside Partnership. In the context of assessment in practical Physical Education at Key Stage 3, the focus here is on verbal question and answer not on written form. Its use is commonly reported for checking understanding and progress, as well as keeping pupils on task, and is often reported as being used to authenticate judgements made through other assessment methods, particularly teacher observation.

At this point in this thesis, the evidence of the reported use of question and answer for assessment purposes and its consistent level of use throughout the period of the present research are simply stated. Later in this chapter, its role in relation to issues of validity and reliability in research question two is examined. Given that teachers use question and answer constantly as part of their teaching and learning strategy, a higher level of agreement might have been anticipated. One theory may be that as teachers do not always plan their questioning in great detail, indeed questions posed are often part of a spontaneous response to the way a particular learning episode develops, their conscious awareness of the extent to which they use questions in assessing learning, may not accurately reflect the reality of their practice. Limitations in the methodology for the present study may then have led to some teachers not reporting question and answer as a specific assessment method.

Peer assessment and self-reflection

The final two assessment methods that are of interest to the present research are peer assessment and self-reflection. Both these methods involve transferring some degree of responsibility for assessment from the teacher to

the learner, thus involving pupils in the assessment process. This is one of the key principles of 'good practice' promoted by Ofsted (2003b) and is a logical progression of the concept of involving pupils in their own learning. The merits of involving pupils in their own assessment are consistently reasoned in assessment research (Black and Wiliam, 1998a; Harlen, 2004a; Black et al., 2003). In short, in this interpretation, assessment is a process in which pupils are actively engaged to inform their own learning. This contrasts with the view of assessment as a process that is done to the pupils by the teacher, in which the learners are the passive receivers of the teacher's assessment decisions. This, therefore, discourages sole reliance on teacher observation, with the teacher passing judgement on the pupils' attainment, and moves to a process where the pupils are part of a partnership, albeit one not of equals, in the assessment of their learning.

This view of assessment would lead pupils to be involved in self-assessment and also peer assessment, for it is argued that through assessing the work of others, pupils will be able to impact on their own learning and attainment. In one study, Tanner and Jones (1994) reported that pupils who had been engaged in discussing and assessing the work of their peers were more successful in conducting their own self-assessments. This was attributed to the pupils developing a clearer understanding of what was required and a heightened ability to reflect back on their own work. Whilst the term AfL, has developed since the Tanner and Jones study was reported, it is possible to see in their findings, early indications of the potential value of the current AfL strategies. Black et al. (2003) found evidence to substantiate these results.

From the data, collected for the present study, there is evidence that the use of both self-reflection and peer assessment have increased across the schools in the Riverside Partnership between 2000 and 2006. At this point self-reflection and peer assessment in the Riverside Partnership are considered separately. However, later in this chapter common issues in relation to the reported practice of both are examined together.

Self-reflection

There is evidence that the use of self-reflection as an assessment method has increased significantly through out this study, from 17% agreement in 2000, to 40% agreement in 2005 and 48% agreement in 2006 (Table 4.4). Indeed, it was not even reported on in the initial data collected in Method A, as it was so rarely seen in the practice of the physical education teachers in the case study partnership schools. This may be explained in that pupils' self-reflection has been promoted widely, (see Ofsted, 2003b) and is a key feature of the SNS.

In light of the earlier discussion of the political impact of Ofsted inspections on the assessment practice of PE teachers, it may be argued that this change has been brought about by the work of the SNS and of Ofsted (2003b). They go on to suggest that the very best practice is seen in schools where opportunities for self-assessment are part of a planned assessment strategy. However, the data collected for this research would suggest that whilst there is clear evidence of an increase in the opportunities for self-assessment, between 2000 and 2006, the evidence does not support the view that a planned strategic approach is in place across all schools in the Riverside Partnership.

Involving pupils in their own assessment, as part of a strategy of involving them in their own learning, is a central tenet of formative assessment or assessment for learning (Black and Wiliam, 1998b; Mansell, James and the ARG, 2009; Ofsted, 2003a) which has been discussed in Chapter Two of this thesis. As previously stated, there is clear evidence from the data that use of self-assessment has increased through out the period of the present research, with a rise from 17% of schools reporting some use of self assessment in 2000 to 48% in 2006. Whilst this reported increase is seen as positive, it is the nature of such self-assessment that will now be explored.

Researchers in the field, such as Black et al. (2003) agree that for effective self-assessment to take place it should be done by the pupils against known

criteria, written in language that the pupils understand. Black et al. (2003, p.31) go on to link the importance of shared criteria to feedback stating that:

> *...shared criteria represent the framework from which teachers evolved appropriate comments to provide information to learners about achievement and for improvement, with self assessment it formed the framework that helped learners decide how to make judgements about their own work and how to structure or detail their next piece of work.*

This is consistent with the view promoted by Ofsted (2003b, p.3) who report as good assessment practice that:

> *Some departments are providing clearly structured opportunities to ensure that pupils are involved in the assessment process and take some responsibility for assessing their own performance against known and understood criteria.*

The data collected for the present research, provides evidence of a mixed approach to the use of criteria across the schools in the Riverside Partnership by 2006. Whilst a number of schools report that criteria for assessment have been written in pupil-friendly language, based on the National Curriculum Attainment target at Key Stage 3, a variety of ways in which schools share these criteria with the pupils are reported in the data. The most common are commented on below.

One approach is to display criteria on noticeboards around the department and in the changing rooms for the pupils themselves to read, before or after their lessons:

> *The attainment levels, adapted into language that the pupils can understand are displayed around the PE block so that pupils can identify which level they are at, and how to improve to*

achieve the next level. This can motivate pupils, as they understand each of the levels (Wetland School, 2005).

However, a weakness of this approach is that they are not on hand at the time that the assessment is being made, in the context of the lesson being undertaken. Therefore, the pupils can only reflect on them before or after lessons have taken place. In addition, they focus on the overall summative level that needs to be awarded at Key Stage 3 rather than the specific criteria for assessment for the particular task being undertaken. This then limits their use as a 'framework for feedback', between pupil and teacher, as proposed by Black et al. (2003) in ongoing assessment during lessons.

Another approach adopted by a number of schools is to collate these criteria, again based on the National Curriculum End of Key Stage attainment levels, into an assessment booklet or profile for PE:

Self-assessment occurs within the school. The use of colour coded assessment strands worded so that they are easy for the pupils to understand enables the pupils to look at the assessment criteria and decide what level they believe they are at and what level they believe they can reach. To help with this a year 7 PE and Games booklet has been produced in which the pupils record what they have learned and what level they believe they are at (Croft School, 2006).

Whilst it could be argued that the booklet would be more accessible for the pupils within a particular lesson, again the focus is on reaching the final level of attainment to be reported at the end of Key Stage 3 rather than on specific criteria for assessment based on the lesson's learning outcomes. In attempting to make sense of these findings, it is necessary to return briefly to the issue of experience and skill in assessment practice by the PE practitioners and teachers.

Whilst there is clearly evidence in the present research that practice in PE has changed in relation to self-assessment and agreed criteria throughout the period of the study, there is some evidence that some times a product rather than a process approach to addressing these issues has been adopted. To clarify, shared criteria, self-assessment and involving pupils in their learning have indeed been and are still being promoted through the SNS, with specific reference to changes to assessment practice.

The data for the present study suggests that some PE departments or individual PE teachers in schools across the partnership have attempted to implement some of these changes into their own practice. However, with limited guidance specific to PE, where some of the challenges do vary compared to traditional core subjects such as Maths, Science and English, due to the fleeting nature of performance in an essentially practical subject, success has been limited. It may be argued that the 'product' of devising the shared criteria, in pupil friendly language has been devised, but the 'process' of using these to meaningfully engage pupils in their learning has not been as successfully achieved. Therefore, their usefulness in contributing to the learning process is still limited.

In relation to the sharing of criteria, one school reported that the use of these standardised assessment criteria, with child-friendly copies made available to the pupils has increased the objectivity of their assessment.

> *Objectivity is addressed using standardized attainment criteria in practical PE, based on the National Curriculum levels. The teachers are provided with a copy and all pupils are shown child-friendly copies to ensure everyone is clear on what is being assessed and what is required to achieve certain levels (Polefence School, 2005).*

Whilst the issue of objectivity is further explored in relation to Research Question Two later in this chapter, it has been noted here to illustrate the variety of claims made for sharing criteria within the partnership schools.

This in turn serves to illustrate the varied interpretations that exist in the Riverside Partnership schools, in relation to the nature and purpose of shared criteria at Key Stage 3 in PE.

Peer Assessment

The reported use of peer assessment, across the schools in the Riverside Partnership, increased between 2000 and 2005 and showed a very slight decline in 2006 (Table 4.4). Peer assessment was reported in the initial study as the second most used method of assessment after teacher observation, although its usage was significantly lower (see Table 4.2). Given the promotion of peer assessment as part of the SNS, this finding was surprising, as I had expected to see a similar increase in its use as was seen in relation to self-assessment.

The practical nature of PE at Key Stage 3, lends itself to the use of peer assessment. A typical use of such assessment is reported in the following example reported from Goldvalley School in 2006. Here one of the student teachers reports on a Y8 gymnastics lesson, which his mentor taught, and he observed. It is a typical example of the reports of peer assessment collected throughout the study:

> *For lesson two, peer assessment was used throughout; therefore, the pupils were now doing the majority of the assessment. When one group began performing, another group was assigned a pupil each to watch. Each pupil gave the other some feedback and had to tell them what attainment level they thought they should receive. The table was left on the board for the children to refer back to. Pupils had to explain why they gave them that level and any ways they could improve their performance (Goldvalley, 2006).*

This commentary shows that peer assessment undertaken through pupils working together, observing each other and giving each other feedback, is a useful assessment method in the context of a practical PE lesson. In line with the work of Black et al. (2003) there is evidence to suggest that criteria were shared and feedback was used to inform learning, in that the assessing pupil, not only suggested a potential level to be awarded, but also suggested

ways that the pupil being assessed could improve their performance. However, given that this was the final task of the lesson, and the next lesson was a week later, opportunities to use this feedback to inform learning were limited. It is also interesting to note that whilst this was a year 8 class, the pupils were being assessed against the attainment levels from end of Key Stage 3, year 9.

Black et al. (2003) distinguish between simple feedback and feedback that contributes to formative assessment. In short, they suggest that feedback can only be part of formative assessment if criteria are shared with the learners, and that the feedback given relates directly to the shared criteria and then this feedback is used to inform the learning process:

> *...formative assessment is a process, one in which information about learning is evoked and then used to modify the teaching and learning activities.[...] Insistence on a precise criterion does not imply a restricted range of activities. The evidence can be evoked in a wide variety of ways [...] but to identify this process on its own as feedback is a serious misunderstanding, albeit a common one [...] Feedback can only serve learning if it involves both the evoking of evidence and a response to that evidence by using it in some way to improve learning (p.122).*

In the example of peer assessment presented earlier in this discussion (Gold Valley, 2006), the summative attainment levels are being used as criteria for assessment against which feedback is given within the lesson. This raises questions as to the meaningfulness of such feedback in terms of its use in a truly formative way, as defined by Black et al. (2003) and its potential to impact on the pupils' learning. It also is a practice that is in contradiction of the guidance issued for teachers by QCA (1999, p.3) who state:

> *Level descriptions are not designed to be used with individual pieces of work.*

Whilst the use of learning outcomes is further discussed later in this chapter, it is worthy to note that many schools in the Riverside Partnership reported linking assessment criteria to the planned learning outcomes or learning objectives of the lesson. Of these, some reported sharing the learning outcomes with pupils at the start of the lesson, and even revisiting them throughout the lesson to monitor pupils' achievement of them. Some mentioned sharing them with pupils via posters or PE handbooks. However, none specifically reported the sharing of assessment criteria linked to learning outcomes with pupils specifically for the purposes of peer assessment. The only criteria reported in this context were the end of Key Stage 3 attainment levels. This issue is further discussed, in relation to involving pupils in the assessment process, later in this chapter.

Involving pupils in the assessment process

Table 4.8 below shows the changes in involving pupils in the assessment process found in the schools in the Riverside Partnership. In seeking evidence of 'good practice' in the data, my analysis was informed by the principles of good assessment practice in PE promoted by Ofsted (2003b) and the work of Harlen (2004a).

Table 4.8 Involving pupils in the assessment process	2000	2005	2006
Detailed assessment criteria linked to learning goals	45%	56%	71%
Opportunities for pupil peer assessment against known and understood criteria	30%	53%	57%
Opportunities for self-assessment against known and understood criteria	10%	25%	46%
Opportunities to observe and evaluate each others' work to identify areas for improvement	70%	91%	100%
Shared criteria for assessment in language pupils understand and pupils understand assessment criteria and know what they have to do to meet them	20%	41%	64%
Progress and attainment recorded in pupil owned progress file, booklet or record of achievement	30%	31%	46%
Progress and attainment recorded in staff controlled database	10%	38%	64%

Given the earlier discussion regarding the lack of evidence of teachers linking shared assessment criteria to learning goals, specifically in relation to self and peer assessment, the 71% agreement from the data in 2006, that this is in evidence across the schools in the Riverside Partnership, appears contradictory, see Table 4.8 above. However, there is evidence in the data collected for the present research, to suggest that teachers are in fact interpreting the NCPE (2000) attainment levels as learning goals. The following commentary serves to illustrate the way in which attainment levels were being used as assessment criteria in a lesson in which pupils were involved in the assessment process, using both peer and self-assessment:

> *Before they [the pupils] performed, the teacher gave each group time to discuss what they think is needed in their sequence to gain a level 4, 5 or 6. Using the whiteboard, the teacher then used question and answer to create a table of all the criteria the pupils thought were needed to achieve each level. This really*

helped pupils to understand what skills they needed to show to achieve their target level (Goldvalley School, 2006).

Whilst it might be argued that this exemplifies 'good practice' in terms of shared criteria in a process of peer and self-assessment, the issue is the use of summative end of Key Stage 3 levels, rather than lesson learning outcomes, as the basis for the assessment judgements. This practice seems to be quite common across the partnership in 2006. It is exemplified in this extract from Croft School (2006):

Formative assessment occurs all the time. Assessment is ongoing during every lesson, so during lesson the teacher can say who is performing the best and who is struggling. Teachers are constantly thinking about what levels pupils are at, and in most lessons they will have an idea as to what level pupils are at. Summative reviews occur at the end of a module of work. These reviews are done by giving the pupils an end of key stage descriptor (EKSD), which is a level of 1 (being poor) to 8 (being sporting excellence). The use of colour coded assessment strands, worded so that they are easy for the pupils to understand, enables the pupils to look at the assessment criteria and decide what level they believe they are at and what level they believe they can reach.

In this extract, there is some mention of AfL practice, in pupils identifying where they are in their learning and where they need to get. However, how to take the next steps in order to progress there is not indicated. This might be interpreted as the "letter of AfL" rather than the "spirit of AfL" (Marshall and Drummond, 2006). It is also interesting to note that the teachers claim to know the levels of each pupil in every lesson and the apparent use of end of Key Stage attainment levels in every lesson. However, summative attainment levels in Physical Education Key Stage 3 were not designed to be used in this way (QCA, 1999); rather they were intended to be used as 'best-fit' descriptors of a pupil's overall attainment, at the end of a Key Stage.

Whilst interpreted as 'good practice' by the teachers at the time, this example serves to illustrate the later concerns of Frapwell (2010) about teachers using levels in ways that were never intended.

From the present research, there is evidence to suggest that teachers from across the schools in the Riverside Partnership are interpreting summative attainment levels as learning goals. This clearly was not the intention of QCA who authored and developed the National Curriculum. On reflecting on this finding, I would suggest that the PE teachers, possibly as a result of lack of experience, adopted the levels as a means of ensuring they were able to evidence their final summative judgements to a variety of audiences. In this interpretation, linking their lesson assessments so closely to the attainment levels against which they needed to report at the end of Key Stage 3 provides evidence of progress to support their final summative judgments. I would purport that this perceived need to evidence their decisions might be a result of an increased need for internal accountability, for example heads of departments through to head teachers or external accountability including parents and Ofsted.

During the interview, when asked about the use of levels, the teacher from Wetland School (2006) commented that:

> *...pupils are given a level for each activity, by learning strand. These are recorded in a departmental database and a modal value is calculated for each pupil at the end of each year. [...] These are reported to parents and the database is used to show progress when Ofsted comes .*

From the data, there is evidence of progress in all aspects of practice related to involving pupils in the assessment process (see Table 4.8 above). The opportunities for self-assessment against known and understood criteria have increased significantly from 10% agreement in 2000 to 46% in 2006. This should be considered, alongside the opportunities for pupil peer assessment against known and understood criteria, which rose from 30% in

2000 to 57% agreement in 2006. Of particular interest here is that both the peer and the self-assessment were based on shared criteria, against which feedback was given. This relates to the earlier discussion of the need to have such criteria upon which to base feedback in order to inform learning (Black et al., 2003).

There is some evidence in the data to suggest that in some schools across the Riverside Partnership, rather than simply watch someone else perform and randomly comment or make a judgement, pupils are being given criteria, which they use to structure their judgements and feedback (Table 4.12). This is consistent with Ofsted (2003b, p.4) who report that:

> *The most effective departments ensure that pupils have well-structured opportunities to develop their observation and evaluation skills across a key stage.*

Whilst there is evidence from the present study that some schools engage pupils in using shared criteria, this is not in evidence in all schools. However, opportunities to observe and evaluate each other's work to identify areas for improvement are in evidence across the Riverside Partnership, with the highest level of agreement possible being noted in 2006 (100%). This means that every school reported that they regularly engaged pupils in observing each other and identifying areas for improvement. However, consistent with the work of Black et al. (2003), it is not possible to comment on the quality of the judgements made, the feedback given or its role in informing learning and progress. As stated earlier in this chapter, given the practical nature of PE, it is a commonly used teaching and learning strategy to involve pupils in observing their peers and giving feedback to identify what could be improved.

The following vignette from a student teacher's observation of his mentor teaching gymnastics with Y8 boys at Rivermeadow School exemplifies the nature of such practice that was commonly reported throughout the schools, in 2006 in the Riverside Partnership. It is clear that whilst some guidance

has been offered to the pupils to help them to structure their observation and feedback, no clear criteria have been articulated to make this truly formative assessment as conceptualised by Black and Wiliam (2009):

> *Methods of assessment involved pupils assessing each other. This was done through partners assessing each other. Half the class performed the other half observed. The pupils observing were asked to assess their peers, in particular to watch out for clear evidence of matching. When the performances were completed, the teacher asked the pupils to pick out some individuals who clearly demonstrated matching. They then discussed what was good about it and then the individuals were asked to perform again to allow everyone to see what was discussed. It was clear that the teacher was assessing their ability to pick out matching movements, therefore assessing their understanding of what a matching movement is. She was also assessing the pupil's ability to evaluate. This type of assessment provided feedback to the pupils. However, the overall feedback was given to the class rather than individuals (Rivermeadow, 2006).*

This commentary does illustrate how pupils are being engaged in their learning. However, it would seem that within the Riverside Partnership schools, as exemplified in this school, opportunities to use this for assessment purposes are perhaps being missed.

Having reviewed the assessment practice in the schools in the Riverside Partnership, attention is now turned to issues of reliability and validity that are of interest to the present research.

Research Question Two

In what ways do teachers of Physical Education, in the Riverside Partnership, consider the concepts of reliability and validity in their assessment practice at Key Stage 3?

Validity is concerned with the extent to which an assessment measures what it was intended to measure and reliability is concerned with the replicability of an assessment. In the literature review, I accepted Harlen's (2004a, p.7) definition of validity and reliability, with validity being interpreted as:

> *Reliability refers to how accurate the assessment is (as a measurement); that is, if repeated, how far the second result would agree with the first.*
> *Validity refers to how well what is assessed, matches what it is intended to assess.*

In order to examine the extent to which teachers in the present study considered validity and reliability in their assessment practices, it is necessary to begin by examining the purposes of, and approaches to, assessment evidenced in the data.

Purposes of Assessment

This study is concerned with internal assessment undertaken by teachers in PE at Key Stage 3. The term 'teacher assessment' is used to describe both the ongoing everyday assessment, which takes place throughout a key stage and the judgements made by the teacher at the end of a key stage. The use of the term 'teacher assessment' does not determine the assessment methods, as discussed earlier in this chapter in relation to research question one. Neither does it designate the approach to assessment in terms of it being formal or informal, nor does it stipulate the purpose of the assessment, whether formative or summative. Formal assessment can be defined as 'assessment conducted in situations solely for that purpose', whereas

informal assessment is 'assessment conducted while pupils are carrying on normal classroom activities' (Satterly, 1981, p.352).

The influence of formal approaches to assessment in PE has been examined in the literature review, Chapter Two of this thesis. Piotrowski and Capel (2000) suggested that informal assessment in PE was less 'systematic' and did not require clearly identified criteria. This interpretation of informal assessment was common in the PE community, both with researchers and teachers, at the time of starting the present study. It was one that I found very easy to accept as it offered a clear distinction between these two approaches that was easy to understand and a clarity that I was able to promote through my seminars to my students.

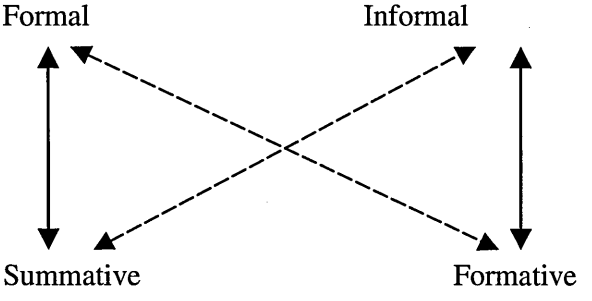
However, as the present study has progressed, informed by the work of Black et al. (2003) in relation to shared criteria and formative purposes of assessment, I began to question this definition of formal assessment in relation to the need for criteria. If criteria are not required, then the logic of the argument is that all assessment for learning can only be undertaken through formal approaches to assessment. Clearly this is a contradiction, and I suspect not what Piotrowski and Capel intended. Therefore, for the purposes of this study, formal assessment is referring to a more explicit, systematic focus on assessment, which is done at a time specifically, set out for assessment purposes, whereas informal assessment is ongoing, less systematic and part of everyday teaching and learning activities.

In the literature review, I accepted the following definition of teacher assessment used for summative purposes:

> *Assessment by teachers for summative purposes means; any activity in which teachers gather evidence in a planned and systematic way about their students learning to draw inferences based on their professional judgement to report achievement at a particular time (Harlen, 2004a, p.1).*

The relationships between the approach to assessment and the purpose of assessment are not mutually exclusive. Figure 4.1 below shows the possible combinations of each approach to assessment with each purpose.

Figure 4.1 Relationships between approaches to assessment and assessment purposes



The model indicates that whilst the most common links are made between formal approaches and summative purposes, and between informal approaches and formative purposes, there is also merit in considering the formative use of formal assessment and the summative use of informal assessment. Examples of each might include, using a video recording of a formal assessment of pupils' practical performance, such as a short sequence in dance, to help pupils to further develop their compositional skills, or using ongoing teacher assessment of pupils' planning skills in games, to inform their overall summative grades for PE. This latter one is of particular interest to the present research and will be returned to later in this chapter, in relation to research question three.

Table 4.9 Assessment Purposes and Approaches

Assessment Approaches	**2000**	**2005**	**2006**
Formal	75%	69%	79%
Informal	80%	88%	100%
Assessment purposes	**2000**	**2005**	**2006**
Summative	80%	81%	93%
Formative	70%	81%	93%

Table 4.9 shows that from the data collected for the present research, there was evidence of both formal and informal approaches to assessment being

used in Physical Education at Key Stage 3, and that such assessment was being used for both summative and formative purposes. Between 2000 and 2005, there is some evidence of a small decrease in the use of formal approaches, which may be accounted for by the developments at national level in promoting an approach to assessment that is embedded as part of ongoing teaching and learning. However, it should be noted that there is then a 10% increase in agreement between 2005 and 2006, with a net result of an overall increase from 75% in 2000 to 79% by 2006. Informal assessment however, has steadily increased in each year of the study from 80% agreement in 2000 to 100% agreement in 2006. This appears to support my contention that the value of embedding assessment as part of the overall teaching and learning process has been recognised by PE teachers in the schools across the Riverside Partnership. This finding also suggests that in the period of the present research, 2000 to 2006, there is some evidence that the tendency towards the use of more formal assessment methods in PE, as discussed in the literature review, is halting. This contention would have been strengthened had a comparable decrease in the use of formal assessment methods been in evidence in the data for 2006. This increase in relation to formal assessment noted in the data for 2006 is intriguing and is worthy of further research.

In all years of the study, the majority of schools report the use of ongoing informal assessment throughout the unit of work, then formal assessment tasks at the end to determine a level of attainment. Whilst this methodology is still very common across the partnership in 2006, there is evidence in some schools that the balance between formal and informal approaches has shifted:

> *Students are continually assessed on a lesson-to-lesson basis, with a formal assessment occurring at the end of each section of study. The results of these ongoing assessments are noted in teachers' planners, and are taken into account when formal levels are awarded (Rivermeadow School, 2006).*

Evidence from the data collected for the present study indicated that informal approaches to assessment included a variety of assessment methods, as is exemplified in this cameo from Pineforest School (2005):

> *Students are provided with several opportunities in each activity to assess their progress through informal peer assessments, written tasks, group assessments and reciprocal work.*

Informal assessment was most commonly reported in the data, as being used to check pupils' learning against the learning outcomes of the lesson, as is exemplified by the commentary cited below:

> *Each lesson is informally assessed by staff through question and answer at the end of each lesson to discover to what extent the learning outcomes have been reached (Polefence School, 2006).*

Formal assessment was reported as being used mostly for final lessons of a unit of work, where the lesson was set aside for assessment purposes, exemplifying Satterly's (1981) definition. This was reported frequently throughout data collection periods, both though Method B in all years and in Method D in 2006, and is exemplified in the comment from County Springs School (2006):

> *At the end of the unit, the pupils are told there will be a formal assessment and that they will be graded on their performance.*

It was also frequently reported that these formal assessments would be judged against criteria from the end of Key Stage 3 attainment levels:

> *All the formal assessments within the department are based upon the NCPE attainment levels (County Springs, 2006).*

and that the level allocated was recorded in a departmental database or pupil-owned profile, as discussed earlier in relation to research question one.

> *The types of assessment used in the PE department to gather evidence of pupil attainment and progress are that the teacher assesses at the end of each unit of work [...]. Every teacher carries out this process for every group they teach [...]. The results are distributed on an Assessment manager database that all PE teachers have access to (Cathedral City High School, 2005).*

Whilst formal assessment was reported for each year of the data, recognition of the limitations in terms of validity and reliability were also noted. To clarify, in the data collected in 2000, it was frequently argued that such one-off final tasks were valid and reliable because they were an 'objective test' of pupils' progress and attainment, and that there was no need for them to be informed by formative teacher assessment. This mirrors the arguments made to justify external testing in the core subjects, (Maths, Science and English) and may be interpreted as PE teachers adopting similar procedures as these subjects as part of a validation of the place of PE on the National Curriculum.

In the 2005 and 2006 data, however, there is evidence that this emphasis is changing; the limitations, of such 'high stakes' (Black et al., 2003; Harlen, 2004b) tasks in terms of their impact on validity and reliability are beginning to be recognised by some respondents. Reasons given for not relying solely on end of unit formal assessments included pupils' having a bad or a good day, so performance observed may not be typical, impact of nerves due to the importance of the assessment, or the teacher simply observing pupils' best or worst performance. The following commentary, from South Pastures School (2006), illustrates this recognition:

The mentor recognised the problems, with end of unit assessment regarding objectivity, validity and reliability, and described several procedures to ensure valid and accurate assessments [...]. The main way to ensure a valid, reliable and objective assessment of a pupil was described to be the use of a range of assessment strategies, conducted in conjunction with each other (South Pastures School, 2006).

It should be noted that this recognition of the need to use with a wider more eclectic range of approaches is widely reported in the data from the schools across the Riverside Partnership in 2006. In many schools, there is evidence in the data to suggest that this understanding has led to changes in teacher assessment practice being implemented. However, the data is inconclusive as to the extent to which this greater understanding impacts on teacher assessment practice in all the schools across the Riverside Partnership.

Assessment purposes

Table 4.10 Assessment purposes

Assessment purposes	2000	2005	2006
Summative	80%	81%	93%
Formative	70%	81%	93%

There is evidence in the data collected for the present study, that assessment practice in PE across the schools in the Riverside Partnership includes both assessments for summative as well as formative purposes. Whilst there is some increase in recognition of both these assessment purposes in 2006, the evidence suggests that they were equally valued in 2005 and 2006. It is worth noting that formative assessment did increase from 70% agreement reported in 2000 to 81% agreement in 2005. This increased recognition of the value of formative assessment since 2000, noted in the data, is consistent with developments at national level, in recognising the value of formative assessment practice. However, given the extent to which this was promoted through the National Key Stage 3 Strategy, a higher level of agreement might possibly have been expected.

With the exception of formative assessment in 2000, there is no noticeable change in the purposes of assessment reported throughout the whole period of the study. This is consistent with the interviews undertaken (Method D) where all mentors interviewed reported undertaking assessment for both purposes:

> *used a mixed strategy of both formative and summative assessment (Teacher, Mansion High School, 2006).*

Assessment is a complex, multifaceted decision-making process, involving teachers in decisions about the methods and approaches to use and about the purposes, that such assessment is intended to address. Having established the assessment methods, purposes of assessment and approaches to assessment, in evidence across the schools in the Riverside Partnership, this section looks at the extent to which teachers considered the concepts of validity and reliability in relation to their decision-making in assessment practice.

Validity and Reliability

Table 4.6 Views expressed Validity and Reliability	2000	2005	2006
Justification of reliance on teachers' professional judgement	70%	19%	29%
Validity considered	55%	84%	86%
Reliability considered	45%	81%	86%

The link between teachers' justification of a reliance solely on their professional judgement and their reported consideration of validity and reliability in evidence in the data, has already been discussed in relation to research question one, earlier in this chapter. This has been detailed in Table 4.6, which for clarity has been reprinted above. In 2000, there was a 55% agreement that teachers across the Riverside Partnership considered issues of validity when making decisions about their assessment practice. This rose to 86% by 2006. In comparison, there was a 45% agreement that teachers

across the Riverside Partnership considered issues of reliability, rising to 86% by 2005. In both cases, this increase is noteworthy, but of particular interest is the rise in consideration of reliability, which has increased, from a lower base to match the increase in consideration of validity.
At the time of conceiving the present research, teachers frequently argued that their professional judgement or experience as a teacher meant that their assessment was valid or reliable, and no further consideration was given to these concepts. This attitude is exemplified in the following comment:

> *My teacher suggested that the validity for assessment came from the professional judgement of his observation and Q and A (Churchenfield, 2000).*

However, the data seems to suggest that some teachers' practice changed during the period of the present study in relation to the extent to which they were aware of the need to consider both reliability and validity in their practice, and an increased reluctance was noted in justifying professional judgement in this way.

Before presenting examples of such practice found in the data, it is worth considering possible reasons for this increase. One that must not be ignored is the impact of the teachers being involved with the PE initial teacher training partnership that is the focus of the present research. As previously discussed, many of the mentors were graduates of the teacher-training provider, which is central to the study. As such, they had all been present at my lectures and seminars. All had attended mentor development events on assessment, specifically on how to improve the quality of their trainees' assessment practice, during the period of the research. Some had even been trainees and had been involved in data collection in the early years of the present research, before becoming mentors as they gained three or more years teaching experience.

When my institution was subject to a full PE subject inspection, 1998 and 2002 – 2003, assessment had been a particular focus. In this context, all

trainees and mentors at that time were involved in very specific training in relation to Ofsted's interpretation of good assessment practice (2003b). Therefore, whilst this is not the only reason for the reported changes in practice, the possible influence on practice of their involvement, with the teacher training and education provider cannot be ignored.

This influence may be accounted for on two levels. One interpretation is that this influence has resulted in a change in assessment practice. However, an alternative interpretation is that mentors are not prepared to be honest in reporting their practice, and thus have skewed their responses to the trainees, in order to appear to have adopted best practice, as articulated by Ofsted and in line with the views I have promoted during such training. However, as previously reported, requiring the trainees to complete their data collection tasks, after a period of block placement in the school, meant that teachers' day to day practice could be observed during this period. This has reduced my doubts to some degree. So having questioned potential weaknesses in the data, it is now possible to examine the reported ways in which consideration of the concepts of validity and reliability are evidenced in the practice of the teachers involved in this research.

In Chapter Two, the work of Harlen (2004a) that reviewed the research evidence of the use of teacher assessment for summative purposes was examined. The findings of this review, included in Appendix Three, have significantly informed the present study, as discussed elsewhere in this report. Whilst data regarding both formative and summative purposes was collected during the whole period for the present research, of particular interest to this study is the extent to which issues of reliability and validity are considered in teacher assessment that is then used in the allocation of summative end of Key Stage 3 attainment levels. Some of these assessments may serve a dual function in that they were conducted as part of a formative (AfL) strategy, but subsequently informed by their knowledge of pupils through such assessments, the teacher used this information to help them determine the pupils summative attainment levels. Thus, formative assessment could be used to contribute to summative purposes.

In the literature review, I accepted Harlen's (2004a, p.1) definition of teacher assessment used for summative purposes:

> *Assessment by teachers for summative purposes means; any activity in which teachers gather evidence in a planned and systematic way about their students learning to draw inferences based on their professional judgement to report achievement at a particular time.*

However, in gathering data about teachers' assessment practice in Riverside Partnership, the evidence of the extent to which 'systematic' approaches were used is confusing. If 'systematic' were interpreted as formal approaches, then this would exclude much of the reported assessment practice that is used to inform the summative end of Key Stage 3 attainment levels. Therefore the focus is moved away from 'systematic' approaches to 'planned' assessment activities, through which the information was used to inform the summative decisions made, as this seemed to represent more accurately the practice reported for the present research, and allows for the summative use of formative assessment practice.

Using the framework for analysis, detailed in Chapter Three, I examined the extent to which the conditions that affect dependability of assessment were in evidence in the PE teachers practice in the Riverside Partnership.

Table 4.11 Conditions that affect dependability	2000	2005	2006
Well developed assessment policy, explicit guidance about the purposes and procedures for assessment	25%	34%	43%
Awareness of potential teacher bias, due to irrelevant factors of behaviour, gender, SEN	15%	34%	43%
Whole school action on assessment, e.g. PPA time	25%	53%	57%
Whole school positive culture for assessment, e.g. shared discussions	35%	69%	86%

As can be seen from Table 4.11 above, in every year of the present study, an increased agreement in all conditions that affect dependability was noted. Harlen (2004a) defined the notion of 'dependability' as a combination of both validity and reliability. Her premise was that these two concepts are not independent of each other, suggesting that 'as one increases the other decreases'. This table therefore represents an interpretation of those findings of Harlen's review (2004a), which were of interest to the present research combined with similar issues identified by Ofsted (2003b).

From the data, there is evidence that the developments in assessment practice in PE have not normally been achieved in isolation, but within a wider whole school culture for improving assessment practice. This supports the evidence of impact of the focus on assessment through SNS. Linked to this, the most notable finding is the increase in terms of a whole school positive culture for assessment from a 35% agreement in 2000 to an 86% agreement in 2006. Evidence that this positive culture has been achieved through some whole school commitment can be seen with the whole school action on assessment, increasing from 25% agreement in 2000 to 57% agreement by 2006. The change in teachers' pay and conditions of service to include guaranteed time during the working week for planning preparation and assessment (PPA time), has affected this finding in 2005 and 2006 data, as it did not exist in 2000. This change in culture and protected time to reflect on assessment practice may have had an impact on

teachers' increased awareness of validity and reliability, as reported in the data .

Ofsted (2003b) and Harlen (2004a) both associate the importance of standardisation and moderation processes for ensuring valid and reliable assessments. As can be seen from Table 4.16 below there is a noticeable increase in evidence in relation to such standardisation and moderation procedures through out the whole period of the present research in the case study schools. When analysing the data here, only evidence in relation to Key Stage 3 Physical Education was included. Any mention of standardisation or moderation procedure in relation to GCSE or A level examination courses was excluded.

The findings from the present research show increased evidence of the notion of 'good practice' determined by Ofsted (2003b) and in the key findings of the Harlen (2004a) review. However, it should be noted that in September 2000 a revised National Curriculum was introduced. This required the use of an 8-point numerical scale when determining pupils' attainment levels at the end of a key stage. This replaced the earlier version, which only required teachers to use a 3-point non-numerical scale, each level of which was open to quite broad interpretation:

> *working towards, achieving or exceeding the expected level of attainment for the key stage (NCPE, 1995, p.20).*

Therefore, in interpreting these changes, it should be noted that there was now a requirement to achieve a greater degree of precision in assessment judgements, as the differences between each level were narrowly defined. This, coupled with the change in assessment culture at whole school level, the provision of time available specifically in relation to assessment, and the increased awareness of 'good practice' in assessment through the SNS, could account for the changes seen in the data between 2000 and 2006.

Table 4.12 Standardisation and moderation Approaches	2000	2005	2006
Shared teacher understanding of NC levels of attainment through opportunities provided for teachers to share 'good practice' in assessment.	35%	66%	89%
Informal discussion between staff to moderate and determine levels.	50%	84%	89%
Formal standardisation through discussion of pupil work to establish criteria for performance at every level.	20%	66%	79%

From my 'insider perspective', with my:

> *a priori intimate knowledge of the community and its members (Hellawell, 2006, p.484)*

it is possible to suggest that following the introduction of the NCPE (2000), and the requirement to assess pupils against more detailed attainment levels of assessment, PE teachers recognised the need to move from a dependence on observation and professional judgement towards more valid and reliable assessment strategies. From this perspective, the increased levels of agreement in relation to standardisation and moderation procedures seen in the data are a reflection of their progress in this ongoing development.

In 2003, Ofsted reported that whilst most PE departments had effective procedures for internal standardisation and moderation of examination course work; it was rarely seen in relation to non-examination PE including Key Stage 3:

> *In exceptional cases, departments use some of these processes [for moderation and standardisation of examination coursework] to moderate the assessment of non-examination work across both key stages (Ofsted, 2003b, p.4).*

In the data, collected for the present study, there was evidence of a significant increase in the value of both informal discussion and formal

standardisation and moderation procedures reported from the schools across the Riverside Partnership, as can be seen in Table 4.12. This growing support for the positive contribution of such procedures is in evidence in each year of the present research, and this finding is consistent with the most recent evidence from Ofsted (2009, p.54), in their review of PE in schools 2005 – 2008:

> *More of the secondary schools used moderation exercises across the department, including moderating judgements about boys and girls.*

However, Ofsted (2009, p.54) go on to state that they found that moderation was not sufficiently widely used:

> *However, when moderation was not applied, this was a lost opportunity to ensure that teachers shared the same standards and high expectations.*

The data from the present study is consistent with this finding. There is compelling evidence in the data for both 2005 and 2006 that the use of moderation and standardisation processes within PE departments increased. However, with only a 79% agreement by 2006, it is clear that some 'opportunities were being lost' in some schools in the Riverside Partnership (Ofsted, 2009).

There was evidence of a range of approaches to moderation and standardisation reported in the data, with practice being reported that is consistent with the findings of Ofsted (2003b and 2009) in terms of both what they term 'good practice' and 'weak practice'. Some common examples are now considered.

The first reflects a very informal approach to moderation and a claim that validity and reliability are ensured based on the experience of the department.

Validity – Experienced department. All teachers have a clear understanding of levels and the requirements to attain these levels. The department can refer back to previous individual to use as a comparison for grades. The members of staff occasionally do some department moderation at Key Stage 3, but this is not considered essential, as they have all worked together for a long time (Churchenfield, 2006).

As can be seen from this commentary, whilst some progress in understanding National Curriculum levels had been achieved, moderation meetings are not considered essential. Validity and consistency in teacher assessment judgements is based on the length of time members of staff have worked together in the department and their knowledge of their pupils and of their subject. Thus, they refer back to previously taught pupils, to inform their decision-making. This is an example of the type of view that was commonly expressed at the time of conceiving the present study in 1998, and which first led to my interest in undertaking research into assessment. That is not to suggest that such a view has entirely disappeared. Indeed in the interview (Method D) the teacher from John Singleton School maintained:

The experience of the teachers within the department ensures that all assessment is reliable (Mentor, John Singleton School, 2006).

The second represents those schools, which during the period of the present research recognised weaknesses in their practice but who had yet to decide on what course of action should be taken to improve. It is difficult to determine from the data, the extent to which the weaknesses were recognised with an intention to improve, or simply recognised as part of the process of being involved in the research, with no real intention to make changes. In the interview with the teacher from Bellsunder School, in 2006 (Method D) she commented:

We have had no CPD for assessment in PE. Probably available if we asked for it but not considered a priority, As far as sharing criteria at Key Stage 3, this is not done at all, going to be prioritised – done at GCSE [...] NC attainment levels are not agreed with other teachers, but we know this is a problem and it is going to be addressed. We have a very small office for five of us so we do talk about kid's performances a lot but we don't moderate with each other. Again, we know it is a problem and when we have time, it will be looked at.

In contrast, evidence of very effective moderation and standardisation processes was also reported in the data. One such example can be seen in the following vignette, which is taken from Cornerstone School, (2006):

To ensure that the attainment grades that teachers give to pupils in the assessment process are valid, reliable and are standard across the school, a standardization process occurs. This standardization process includes a number of sessions where pupils perform in a variety of activities in PE. Each PE teacher then assesses each pupil using the school proforma and gives an overall attainment grade relevant to the National Curriculum levels. The teachers then compare the grades that each of them gives for each pupil and discuss the justification for the grade that they have given. The discussions allow each teacher to justify their decisions, and then ensure that they are assessing to the correct and same criterion and therefore ensure validity, reliability and standardisation across the department.

This practice is consistent with the 'good practice' espoused by Ofsted (2003b) and the findings of Harlen (2004a). Again resorting to my 'a priori intimate knowledge' to inform my commentary, (Merton, 1972) it is interesting to note that both schools are from the same Local Education Authority, and in both departments there are graduates from the initial

teacher-training provider in the Riverside Partnership. In Cornerstone School however, the head of PE at that time had a particular interest in assessment and since the completion of this study was appointed to an assistant headship, with school wide responsibility for assessment.
In the data for 2006, there was some evidence of co-marking as part of the moderation systems in place at Key Stage 3. Co-marking is of particular interest to the present research, as it can not only check validity but also serve the purposes of reliability in terms of the extent to which the second result matches the first:

> *Several teachers conduct assessments of individual pupils. This 'cross assessment' or moderation system ensures the reliable results and makes certain that all teachers are working to the correct criteria, ensuring validity (South Pastures School, 2006).*

This development of more attention to validity and reliability reflects the influence of examination courses in PE, where such systems are part of established practice. The impact of examination practices on the development of practices for non-examination work is a theme running through this thesis (Carroll, 1994; Green, 2008). Returning briefly to the interview with the teacher from Bellsunder School, she commented:

> *new ideas that are generated by staff are usually motivated by GCSE [...] then informally drift down into the lower school.*

Whilst there are many examples of standardisation and moderation reported in the data, one final one from Wetland School has been included here to exemplify the whole departmental approach to assessment that has been noted through the period of the present research. This exemplifies the 'positive culture' for assessment identified by Harlen (2004a), where assessment is constructively discussed:

Every couple of weeks the PE department has a meeting to discuss any issues that may have arisen. Every third meeting, the teachers observe and grade a range of pupils in both ability and age. This is to ensure all teachers are familiar with the grading system and to standardize overall grades. In turn, ensuring all results given for assessment is valid and reliable (Wetland School, 2006).

Six local education authorities (LEA) are represented within the sample of schools in the Riverside Partnership. No mention of specific standardisation and moderation procedures being consistent across any individual LEA was noted in the data collected through Method B. However, the teacher from Wetland School, in Romeston LEA reported that through their role as a Sports college, the partnership development manager was attempting to standardise assessment in PE across the LEA (Method D). At the time of the interview, the focus was on developing an assessment package for use at Key Stage 1 and 2 in primary schools. This was the only indication in any of the data collected of any attempt at consistency in practice being attempted across schools in one LEA. In all other schools involved in the research, only whole school or departmental strategies for improving assessment practice were reported. The reasons for this are of interest to me but beyond the scope of the present research. Within Wetland School, the teacher also reported, "collating material from other PE departments, to see what they did", when she set about the task of developing shared criteria. This view that something could be learned from working with other schools was also articulated in one other school, from a different LEA but also a Sports college. In the data collected in 2005 about this school (Method B) concern was expressed about the lack of validity and reliability in teacher assessment in PE:

One area, which is an issue within schools' PE levels, is that some schools, at the end of year 9 just take a pupil's best mark and put that as their overall grade. Validity here must be questioned (Rivermeadow School, 2005).

This school proposed setting up moderation and standardisation meetings with other schools in the area. However, whilst a similar proposal also appeared in the 2006 data, collected at this school, it had yet to be implemented.

> *It is also a future plan for all local schools to have a moderation of assessment day, bringing in a mix of staff and a mix of pupils, help all staff get on the same wavelength in terms of what performance requires which grade (Rivermeadow, 2005).*

With these two exceptions, all other references to standardisation and moderation were at departmental level only. In mentioning that both of these schools had Sports college status by 2005, it should be noted that at the time of first conceiving this study (1998) the development of Sports Colleges, through the Governments specialist schools programme (DfES1997) was in its infancy. Therefore, as at that time only the first cohort of schools were going through their application process, I could not set out to measure their potential impact. During the lifetime of the present study, many of the schools in the Riverside Partnership were successful in gaining Sports College Status. From my 'insider perspective', (Hellawell 2006, p.488) it has taken on average 4 years from first deciding to apply for Sports College Status to the completion of related building works following successful application. In 2000 the Department for Culture Media and Sport published a white paper which set out the Labour government's vision for sport in the UK, entitled, 'A Sporting Future for All'. Included in the document was mention of the role that Sports Colleges could play in achieving their plans, where it was stated:

> *Our Specialist Sports Colleges are at the forefront of developments in school PE and sport. All of them work with other schools to share their expertise, resources and 'good practice', so that locally there is a 'family of schools' working together to provide training and support for teachers in*

secondary and primary schools, and to maximise the opportunities available for all pupils (DCMS, 2000, p.30)

Given their Sports College Status, and therefore the involvement of both schools in 'sharing their expertise, resources and good practice' (DCMS 2000, p.30) it may be that these two schools had established a culture of working in partnership with local schools and that this extends to all aspects of practice including assessment. However, it is open to question as to why a similar interest in working with other schools, was not reported in the data collected from other Sports Colleges. One possible reason may be the year in which they were successful in their application and the time scales to complete their construction. Thus, at the time that the data was collected for the present study, they were all at different stages in their development.

Teachers' constructs of PE and dependability of their assessment practice.

Having considered the evidence from the data collected for the present research in relation to issues of dependability, (Harlen, 2004a) at both whole school and departmental level, attention is now drawn to the teachers' conceptualisation of the construct of Physical Education. How this reflects their values and how this may have an impact on the dependability of their assessment practice in PE is an important theme for the present research.

As previously discussed in Chapter Two of this thesis, PE is a complex concept, about which there exists no universally agreed definition, thus the construct of PE, and what constitutes competence in PE may vary between teachers. In Chapter Two, I suggested that whilst there exists a general agreement that PE is about the development of 'physically educated pupils.' there is significant variation about what this notion of being physically educated actually means. I suggested that it is best understood as a continuum of views. At one end are those who link it closely to sport and see the purpose of PE is to educate the pupils in terms of the knowledge and skills required to engage with the prevailing national and international culture of sport (Alderson and Crutchley, 1990). This view has gained

credence during the period of the present research in light of the successful bid by London to host the 2012 Olympiad. At the other end are those who advocate that the place of PE on the school curriculum is justified in terms of its capacity to educate pupils through physical contexts, where PE is primarily valued as a process of learning, where the context is primarily physical, (Murdoch, 1990; Penney, 2000; Whitehead, 2007; AfPE, 2008; Morley and Bailey, 2006). This view is promoted most strongly through the professional association representing PE (AfPE). As with all spectra, there are a myriad of views located in between these two extremes.

In reflecting on issues of validity and reliability in relation to assessment practice, it is important to recognise the role of teachers' own perceptions of PE within this broad spectrum and how this informs their values and in turn how these impact on their practice. As previously discussed, for many PE is primarily a performance-based subject and those attracted to Physical Education teaching are usually highly competent performers in at least one sport. Using my 'a priori intimate knowledge' (Merton, 1972), gained through a 25 year teaching career, 18 of which have been spent in secondary initial teacher training in PE, most applicants to PE initial teacher training courses come from a games background, with many having achieved national or international sporting honours in at least one. The reasons for this dominance of games are not the concern of the present research, other than to suggest that this does underpin the inextricable links between PE and sport (Green, 2008). Given that most people attracted to PE as subject knowledge experts are competent performers, whilst one might recognise the value of planning and evaluating as well as performing, there is limited evidence to suggest that applicants are attracted to becoming PE teachers because they excel in planning and evaluating! Most excel in performing and this impacts on what they value in terms of competence in PE, which may be reflected when assessing their pupils, leading to an overemphasis on performance.

The NCPE (2000) that is current though the period of the present study reflects an educational interpretation of PE; that of education through the

physical, rather than an interpretation of PE as sport. The pupils are required to be engaged in planning and evaluating as well as performing. The four strands of learning are at the heart of the curriculum and as previously stated, teachers are required to reflect on a pupil's ability in all four to reach an overall 'best-fit' attainment level by the end of each Key Stage. However, if one takes into account teachers' own backgrounds, as highly competent performers, this can lead to an over-valuing of performance rather than planning or evaluating. Thus the pupil who can perform skilfully, but lacks planning and evaluation skills, may be over graded, whereas the pupil who is outstanding in these areas, but lacks the physical skilfulness to execute high level performance is often under graded. If one accepts the premise of this broad interpretation of the construct of PE, which varies by individual teacher, then the difficulties in reaching valid and reliable judgements in PE at Key Stage 3 are exposed. This is heightened when a teacher's personal construct of PE is at odds with the prevailing National Curriculum.

In the data collected for the present research, there was some evidence of this tension. This can be seen in the following extract from the student and mentor interview, (Method B) from Pineforest School (2005):

> *What is being assessed? The assessment criteria at the school are not totally objective and reliable, which is illustrated in the assessment marks given by two different teachers to one class. Furthermore, the assessment concentrates more on the performing ability of the pupils instead of the planning and evaluating aspects. The school suggests that this is more to do with the constraints on time and the feasibility of the assessment in question for example assessing pupils using video recorders requires time and equipment (Pineforest, 2005).*

In the data collected for the present research, there was evidence that all four strands of the National Curriculum are being assessed. In some schools, peer assessment is used to address the evaluation and improving learning strand

of NCPE (2000). This was evidenced in Croft School in 2006, where it was noted:

> *Peer observation and feedback is used, especially within Dance. This links to the evaluating and improving aspect of the National Curriculum. Therefore, with peer evaluation pupils look to see what is good about a performance and what could be improved.*

However, performance was the most commonly reported. This was noted in the data, particularly in relation to the use of teacher observation or video assessment:

> *The assessment is mainly completed through teacher observations and focuses mainly on performance. The teacher uses a criteria sheet, which is specific, to determine the level of the pupils (Pineforest, 2005).*

Given the requirement to reach an overall 'best-fit' attainment level, which reflected all four strands of learning, it was frequently reported in the data that pupils' ability in 'evaluating and improving' was assessed through the units of work in gymnastics or dance, as exemplified through the following commentary:

> *There are times when the pupils assess each other and provide feedback [...] e.g. in gymnastics, where they have to complete evaluation sheets about the performances of other groups (Churchenfield, 2000).*

In interpreting this finding, it should be noted that in addition to the dominance of games, as discussed earlier in this chapter, many PE teachers have a limited personal experience of dance or gymnastics; for many these are new areas of activity to which they are introduced at college as part of their initial teacher training courses. This, therefore, means they have a

lower skill level in dance or gymnastics. This may affect their practice in two ways. One is that they focus on the development of the pupils' compositional skills in the activity rather than developing their practical performance skills, hence the evidence in the data collected for the present research that, through gymnastics and dance, the evaluating and improving strand, and selecting and applying strand of learning in Physical Education are assessed. Secondly, where acquiring and developing strands are assessed in these activities, there is evidence to suggest that they tend to over grade pupils' performance in these areas of activity if the pupils are at a performance level higher than they themselves have achieved. Both of these issues raised can have an impact on the validity and reliability of teachers' assessments in PE.

The issues linked to validity and reliability raised in this debate of teachers as subject knowledge experts reinforces the need to develop effective moderation, and standardisation processes in order improve the dependability of using ongoing teacher assessment for summative purposes. It also has implications for initial teacher training providers in terms of recruiting people from a wider background and developing their understanding of the complexity of the construct of PE and the need to recognise their own values and limitations and how these may impact on their practice.

This chapter has reflected on the assessment methods, approaches and purposes of assessment in evidence in the data collected for the present research. It has also examined the ways in which issues of validity and reliability are considered at whole school and departmental level, in the schools in the Riverside Partnership, and the impact that teachers' personal concepts of PE have on their values and therefore their impact on their assessment practice.

I shall now offer an interpretation of the data in relation to the final research question.

Research Question Three

How do teachers of Physical Education, in the Riverside Partnership, make 'best-fit' judgements, as required by National Curriculum 2000, to decide end of Key Stage 3 summative attainment levels, which are reported to parents?

Assessment practice reported in summative grading of NCPE Key Stage 3 attainment

Table 4.13 below provides a summary of assessment practice, noted in the data collected through Method B, which teachers in the Riverside Partnership schools use to inform their grading of pupils, in order to reach an overall attainment level at the end of Key Stage 3 PE, as required by the NCPE (2000).

Table 4.13 Summative grading of NCPE Key Stage 3 attainment	2000	2005	2006
Formal Levelling against NCPE	40%	91%	100%
End of unit assessment used cumulatively to determine achievement against NC levels of attainment	45%	78%	89%
Progressive levels of attainment applied	30%	53%	75%
Levels recorded + / - to show subtle differences between pupils	0%	19%	25%
Planning and assessment linked to NC programme of study	45%	84%	93%

Across the schools in the case study partnership, increased levels of agreement in all aspects examined were noted in each year of the present research. This suggests that practice in general across the Riverside Partnership has developed through out the period of the present study. A significant increase in the use of formal levelling occurred between 2000, (40% agreement) and 2005 (91% agreement), finally reaching 100% agreement in 2006. This may be explained in relation to the 'roll out' approach to the introduction of the revised NCPE (2000) with its 8-point numerical scale. As discussed earlier in this chapter, prior to this is a non-

numerical three-point scale was used. Teachers are only required to use these levels at the end of a key stage, thus they were first used with the cohort of pupils who reached the end of Key Stage 3 in 2003.

There is evidence in the data that this requirement to assess using the National Curriculum levels also impacted on other aspects of teachers' practice, with a greater level of agreement noted in 2006, (93% from 45% reported in 2000) that planning and assessment are linked to National Curriculum Programme of Study. This finding suggests that teachers across the partnership are mapping their planning and assessments more closely to the NCPE (2000) in order to facilitate the summative end of Key Stage 3 assessment process. This has resulted in changes to their planning as well as their assessment practice. This is consistent with the findings of Coladarci (1986) (cited by Harlen, 2004a, p.3) that:

> *Teachers' judgements of students' performance are likely to be more accurate in aspects more thoroughly covered in their teaching.*

A higher percentage agreement with the use of progressive levels of attainment is noted from 30% agreement in 2000 to 75% agreement in 2006, which again related directly to the use of an 8-point numerical scale, and the Ofsted (2003b) notion of 'good practice'. This reflected the approach taken by many schools in the Riverside Partnership of dividing the attainment levels into expected levels for each year, thus in year 7 pupils might expect to be level 3 - 4, by year 8 level 4 – 5, and by year 9 level 5 – 6. On analysing the data from, 2005, it was noted that some schools were starting to report the use of 'sub-levels', that is levels being differentiated to show subtle differences between pupils at the same level, which had not been reported at all in the data from 2000. Albeit that the level of agreement noted was small, the fact that it increased in 2006 is worth reflection.

The most commonly reported method in the data, was to break each numerical level into a, b and c. Thus, level 5a represents a higher attainment

level than level 5b. This practice, that is a requirement in the core subjects such as Maths English and Science, is not a requirement in PE. However, influenced by Ofsted (2003b) and the increased pressure for accountability in schools, where pupil progress has to be evidenced and based on data, some PE teachers, within the schools in the Riverside Partnership have begun to adopt this practice. In addition, allocating a range of levels to particular years may also offer an insight as to why some schools saw the need to differentiate within a level, in that it allows a pupil's progress to be seen within one particular year. For example, if a pupil begins year 7 at level 3c, and reaches 4c then it is argued that their progress is evident. In the interviews conducted in 2006, (Method D) this use of sub-levels was explored in the two schools in the interview sample of six, who had reported their use.

During my interview with the PE mentor at Wetland School, who was also the departmental Assessment Coordinator, the process for recording attainment in PE was clarified, as follows:

> *The pupils are assessed by strand [NCPE 2000] in each activity in every module using sub-levels. These are then entered into the department assessment recording system. At the end of the year, all scores are aggregated by learning strand not by activity. This means we can see how each student is doing in each of the strands. It means those who are not so good at an activity, for example gym, can still do OK in evaluating and improving or fitness and health. [...] so, 4 levels go on the reports home[...] all scores continue to be aggregated until end of Y9. End of Key Stage 3 reports to parents use overall aggregate scores from Y7 –Y9 (PE mentor, Wetland School, 2006).*

In this interview with the teacher from Wetland School, it is clear that these sub-levels were used with the intention of evidencing and tracking pupils' progress, with a departmental database. This database was quite complex, in that pupils received levels from each of the strands of learning identified by

the National Curriculum for each of the units of work. The modal grade for each learning strand was then calculated, and was used to report on each learning strand in the year 7, 8 and 9 reports. Initially, this approach appeared to be able to reflect the full profile of pupils' attainment, in that each learning strand was assessed in each unit of work. However, the modal grades for each year were combined in order to reach the final grades reported to parents at the end of Key Stage 3. The inclusion of the year 7 and 8 scores in the final year 9 calculations, as well as the year 7 scores in the year 8 calculations meant that pupils' progress was masked, in that lower scores from the earlier years were lowering the final results. Thus, ultimately what had been designed to evidence progress was actually seen to hide progress due to the complexity of the methodology of calculation.

In the interview with the teacher from Mansion High School the only other school in the interview sample who reported the use of sub-levels, the teacher stated:

> *We use sub-levels, a, b, c, so the pupils can clearly see their progress. We also use them in the lessons, so they know what they need to do to get a higher grade. We find these help to motivate the children, who like to have a 'score'. If we did not and a child had improved, but not enough to move, say, from a 4 to a 5 then they would lose heart (Teacher, Mansion High School, 2006).*

As can be seen in the extract from the interview with the teacher from School, over the period of the present study there was evidence that some of the schools were starting to use levels throughout the Key Stage and not just at the end:

> *The attainment levels are displayed around the PE block so that pupils can identify which level they are at, and how to improve to achieve the next level. This can motivate pupils, as they understand each of the levels (Wetland School, 2005).*

From these accounts, it can be seen that not only were levels being used as criteria against which to assess learning, (not just as final summative grades) but also the idea of linking levels to pupils' motivation is also emerging. These extracts, from the interviews with Mansion High and Wetland School are included here to illustrate the ways that many of the schools across the Riverside Partnership report using levels, both formatively and to inform their summative decisions.

Linking back to the discussion presented in relation to research question 2, the following commentary from Croft School, (2006) exemplifies how some schools regarded the use of these levels in terms of helping to improve the reliability of their teacher assessment in PE:

> *Reliability – The End of Key Stage levels were re-written and every member of staff was given a copy, so they now all have the same assessment criteria to assess the pupils and the end of each unit of work. Therefore the pupils should get the correct level, they should get the same level regardless to which teacher assesses them (Croft School, 2006).*

This link to reliability and the use of levels as criteria have been discussed earlier in this chapter, as has the practice of formal assessment undertaken for summative purposes at the end of a unit of work. However, this commentary also illustrates how this practice developed through the period of the present study, with the allocation of a National Curriculum attainment level, being awarded as part of this process. Given that on average a unit of work in PE lasts approximately half a term, (6 weeks) this is a very different use of the levels from what was intended by the National Curriculum authors.

The 'best-fit' approach

The National Curriculum 2000 requires teachers to use a 'best-fit' approach to deciding on pupils' summative attainment levels at the end of each key stage. The statutory advice for determining a level for the various subjects is to apply a 'best-fit' notion which:

> *is based on a knowledge of how the pupil performs across a range of contexts, takes into account strengths and weaknesses of the pupil's performance and is checked against adjacent level descriptions to ensure that the level awarded is the closest match to the child's performance in each attainment target (QCA/DFEE, 1998, p.1).*

PE specific guidance, published by QCA in 1999, reinforced the 'best-fit' approach and that level descriptions were only intended for End of Key Stage use:

> *Level descriptions are designed for End of Key Stage use only. Teachers will determine which level description best-fits a pupil's performance (QCA, 1999, p.5).*

This concept of 'best-fit' was first introduced in 1996. In investigating the practice of the PE teachers in the Riverside Partnership in 2006, my work was informed by the findings of Gipps and Clarke (1996, 1997). These studies, funded by SCAA investigated how primary teachers and secondary teachers in Maths, Science and English make Teacher Assessment judgements. The key findings of these studies are presented in Chapter Two.

This final research question was formulated as the present research progressed. In March 2006, when analysing the data collected for Methods B and D, I found that whilst I had collected data that gave insight into teacher assessment practice during the research period, the evidence was inconclusive as to how they finally used this information to make their

'best-fit' judgements at the end of Key Stage 3 in PE. In order to gain this understanding I decided to simply ask them! I constructed an email questionnaire. In using this, I sought to find out how teachers of PE in the present study were interpreting this notion of 'best-fit', how they were using it in their summative end of Key Stage 3 decision-making, and what evidence from teacher assessments they were using to inform this process. Using similar headings to those used by Gipps and Clarke, (1996; 1997) enabled me to draw comparisons with the practice of the PE teachers and the teachers of Maths, Science and English in secondary schools, and with primary teachers who were the focus of their research. Given that there is no requirement for external testing in PE, this heading was interpreted as 'formal end of unit assessments' as discussed earlier in this chapter. Having analysed the data, as detailed in Chapter Three of this thesis, a number of tables were compiled for ease of analysis.

Table 4.14 How do you make 'best-fit' judgements?		
By making generalised 'best-fit' judgements	20	100%
By splitting the level descriptors (e.g. by creating separate statements and counting half or more as attaining level	14	70%
By identifying key aspects of level descriptions	14	70%
By using 'best-fit' judgements in relation to children's portfolios of practical performance	13	65%
	N=20	

From Table 4.14 above it is immediately apparent that all of the PE teachers in the sample of 20 used generalised 'best-fit' judgements to determine pupils' levels of attainment. This generalised approach was more consistent with the practice of the primary, than secondary teachers, as reported by Gipps and Clarke (1996) although it should be noted that even at primary level it was not universally used (71.7% in Y2 and 76.1% in Y6). However, they did report that it was the most commonly used across their sample.

Table 4.15 How do you make 'best-fit' judgements? Ranked in order of preference.	Total
By making general 'best-fit' judgements	47
By identifying key aspects of level descriptions	67
By using 'best-fit' judgements in relation to children's portfolios	69
By splitting the level descriptors (e.g. by creating separate statements and counting half or more as attaining level	70

Following mathematical manipulation, as detailed in Chapter Three, the data was ranked in order of preference within the teachers' practice (see Table 4.15). Again, it clearly shows that the PE teachers, to all other approaches, preferred a generalised 'best-fit' approach. This provides compelling evidence that by 2006, the QCA, (formerly SCAA) 'best-fit' approach had become a common feature of PE teachers' practice in the schools across the Riverside Partnership.

By presenting the data in Table 4.14 and Table 4.15, it is possible to see not only the percentage of teachers agreeing with each statement, but also the statements in terms of preference in the teachers' practice. Thus whilst 70% (14) teachers reported that splitting the level descriptors was part of their practice, when ranked in order of preference this was seen to be the least preferred method. This may suggest that whilst there was evidence of teachers across the Riverside Partnership breaking down the level descriptors into pupil friendly statements, as discussed earlier in this chapter, the evidence was less compelling that the results of ongoing assessments in relation to these statements were being used to inform final summative grading. This finding was consistent with the study by Gipps and Clarke (1996) a decade earlier that found that splitting levels into separate statements was also the least used approach across all the groups in their sample.

70% (14) of the PE teachers in my own sample did report that they identified key aspects of the level descriptors, (a pupil must be able to x, y, and z); in order to reach a particular level and this approach was ranked

second in terms of popularity. The use of children's portfolios of practical performance links to the evidence of ongoing assessment practice found elsewhere in the data collected for the present research and discussed earlier in this chapter, giving evidence that many schools across the Riverside Partnership claim to be using these portfolios to inform their summative judgements. However, given that the overall score difference between those statements ranked 2 – 4, was so low, it is impossible to draw any meaningful conclusion about preferred practice other than a clear preference for a generalised 'best-fit' approach.

Having identified the preference for 'best-fit', I then analysed the data to see how this concept was being interpreted; the results are presented in Tables 4.16 and 4.17.

Table 4.16 How do you interpret 'best-fit'	Total N=20	%
The level description which overall describes the child's attainment better than the one above or below	18	90%
Must achieve important aspects of a level description	16	80%
Intuition	12	60%
Must achieve 75% or more of the statements in the level description	10	50%
Must achieve almost 100% or 100% of the statements in the level description	4	20%
Must achieve 50% or more of the statements in the level description	2	10%

As was found in the earlier work of Clarke and Gipps (1997) with primary teachers, the most common interpretation of a general 'best-fit' judgement was to decide which level best describes a pupil's attainment better than the level above or below, with 90% (18) of the teachers agreeing to this statement. However, this statement is in itself problematic, as it does not actually elucidate the teachers' decision-making process in terms of how they reach the judgement as to which level is more appropriate. Clarke and Gipps (1997) had considered this problematic but included the statement at the express request of SCAA, who was funding their research. As a result, it was also included in the present study. Not only was it the most common

interpretation of 'best-fit', but also it was also the most highly ranked by the teachers, as can be seen from Table 4.17 below.

Table 4.17 How do you interpret 'best-fit?' Ranked in order of preference.	Total Scored
The level description which overall describes the child's attainment better than the one above or below	38
Must achieve important aspects of a level description	60
Intuition	97
Must achieve 75% or more of the statements in the level description	101
Must achieve almost 100% or 100% of the statements in the level description	141
Must achieve 50% or more of the statements in the level description	146

Table 4.16 and Table 4.17 offer some insight into the teachers' decision making. It is interesting to note that the frequency of reporting of each aspect of practice, corresponded with their reported rank order, thus achieving important aspects of a level description is both commonly reported, 80% (16) and highly ranked, (2^{nd}). Intuition is also reported by a majority of teachers, 60% (12).

Whilst 50% (10) of the teachers required pupils to achieve 75% of the statements identified, some required achievement of 100%, (4, 20%), whereas others only required them to achieve 50% of the statements, (2, 10%). Whilst the numbers are small, this does raise questions in terms of consistency in teachers' practice. This is important in the present research for two reasons. Firstly, it questions the validity and reliability of the summative attainment levels awarded to pupils by different teachers in the Riverside Partnership schools. Secondly, it suggests that the trainee teachers will be receiving conflicting guidance as to how to interpret 'best-fit' and therefore will expose them to inconsistent practice, which will affect the quality of their training. For example, a trainee could spend one placement in a school that only requires a pupil to achieve 50% of the statements to achieve a level 5, whereas in their second placement pupils may be required to achieve 100% in order to attain the same level. This also strengthens the

need for robust standardisation and moderation processes to be developed not only within each PE department but also between schools. As discussed earlier in this chapter the data for the present study suggests that there is some 'good practice' in some schools in the Riverside Partnership, in relation to the former. However, there is compelling evidence that within the Riverside Partnership schools interschool standardisation and moderation processes have not been developed in PE at Key Stage 3 between 2000 and 2006.

Evidence used to decide teacher assessment levels in Physical Education at the end of Key Stage 3

Table 4.18 presents the findings of the final area of interest examined through Method C, in which teachers were asked about the types of evidence they used to inform their decision-making at the end of Key Stage 3.

Table 4.18 Evidence used to decide teacher assessment levels in Physical Education at the end of Key Stage 3	Total N=20	%
Professional judgements based on knowledge of the child	20	100%
Discussion / moderation with colleagues in school	18	90%
Level descriptions used as check lists	18	90%
Children's work	16	80%
Ongoing and termly or half termly tests either in house or commercial	14	70%
Jottings of ongoing assessments: achievements made, help needed etc (weekly / daily)?	12	60%
Discussion with the children	12	60%
Observational notes	8	40%
School portfolios for PE	4	20%
Marking comments	1	5%

It can be seen that by the end of the period of the present research, 2006, 100% of teachers (20) reported using their professional judgement based on their knowledge of their pupils to decide teacher assessment levels in Physical Education at the end of Key Stage 3.

Given this stated preference for 'professional judgement', it may appear at first that little had changed since the study started in 2000. However, the increased evidence of teachers discussing pupils' attainment with colleagues, both informally and as part of a moderation or standardisation process, using the level descriptors and undertaking the ongoing assessments of pupils' work support the findings in the data collected through Method B and D. These findings suggest a heightened awareness on the part of the teachers, of the need to consider issues of reliability and validity in order to increase dependability in relation to their summative end of Key Stage 3 assessment practice.

In this chapter, the data collected for the present study has been analysed and interpreted in relation to each of the three research questions. Chapter Five will consider the key findings, conclusions, and implications of the research.

Chapter Five: Conclusions and Implications

In Chapter Four, the data collected for the present research has been analysed and interpreted in relation to each of the research questions. In this chapter, the key findings of the study have been summarised and the main conclusions and their implications for policy and practice in Riverside Partnership are presented. Finally, it offers areas of interest for a future research agenda in relation to assessment practice in PE.

Summary of the main research findings

The PE teachers in the schools in the Riverside Partnership use a general 'best-fit' approach to determine pupils' summative attainment levels at the end of Key Stage 3. The most common interpretation of 'best-fit' is the level description which overall describes the child's attainment better than the one above or below. However, there is some inconsistency across the partnership in the ways in which teachers are interpreting what a child needs to do to evidence their achievement of a particular level. Where this exists, it has an impact on the validity and reliability of teachers End of Key Stage attainment judgements.

There is evidence of a link between using a wider variety of assessment methods, (beyond teacher observation and question and answer), and decreased justification of a dependence solely on teachers' professional judgement. Thus, although teacher observation is still widely used in 2006 in the schools in the Riverside Partnership, it is supplemented by the use of other assessment methods, which in turn strengthen the dependability of teachers' professional judgements.

The use of video and digital cameras has increased during the period of the present research. However, whilst some schools are using such technology very effectively to support their assessment and moderation processes for PE at Key Stage 3, its use is not as widespread through the Riverside Partnership as had been anticipated.

By 2006, there is less emphasis on using 'one-off' final assessments to determine summative levels of achievement. In 2000, it was frequently argued that such one-off final tasks were valid and reliable because they were an objective test of pupils' progress and attainment. However, by 2006 the limitations of one off tasks in terms of their impact on validity and reliability were being recognised. Whilst many schools do still use final one-off summative assessments to inform final End of Key Stage attainment levels, these are also informed by ongoing formative assessments throughout the key stage to give a more rounded judgement on pupils' attainment.

Though not always mentioned specifically by name, there is evidence of 'Assessment for learning' (Black and Wiliam 1998a) principles being adopted in many schools in the partnership by 2006. Practice noted includes:

- Feedback to inform learning and progress
- Shared criteria for assessment in language pupils understand, including sheets, displays and pupil progress files
- Question and answer to check understanding and inform future learning
- Peer and self-assessment opportunities

There is evidence of development in involving the pupils in the assessment process. In 2000, no schools reported sharing the assessment criteria with the pupils in language that pupils understand. However, by 2006 there is evidence of pupils being made aware of the criteria against which they are being assessed, and in schools with a strong awareness of current assessment thinking, there is an emphasis on ensuring that the pupils know what they have to do to meet these criteria. This emphasis most commonly takes the form of displays around the PE department and sharing criteria within lessons. In a small number of schools, pupil portfolios have also been developed.

By 2006, there is some evidence of teachers developing their use of peer assessment. However, whilst there is significant evidence of peer evaluation and feedback being used in teachers' ongoing teaching and learning strategies, opportunities to move from simple feedback into formative assessment for learning are being missed. In general, feedback used is linked to further development rather than being simple praise or criticism. However, when teachers do not formalise the criteria against which the peer feedback should be given, its use in informing pupil progress, and therefore its potential as assessment for learning, is not being realised in all schools in Riverside Partnership (Black et al., 2003).

Ofsted (2003b) suggest that the very best practice is seen in schools where opportunities for self-assessment are part of a planned assessment strategy. However, the data collected for the present research would suggest that whilst there is clear evidence of an increase in the opportunities for self-assessment between 2000 and 2006, the evidence does not support the view that a planned strategic approach is in place across all schools in the Riverside Partnership.

Whilst there is clear evidence in the present research that practice in PE has changed in relation to self-assessment and sharing agreed criteria with the pupils throughout the period of the study, in some schools, a product rather than a process approach to addressing these issues has been adopted. There are examples of 'good practice' where schools have devised a number of products to share learning criteria with pupils, including posters and displays around the departments, PE handbooks and progress files. However, the research approach adopted, did not allow me to accurately evaluate the processes by which these were used to meaningfully engage pupils in their learning and assessment.

There is evidence from the present research to conclude that a minority of teachers regarded the need to record pupils' progress as an opportunity to engage their pupils in self-reflection. The majority, however, regarded it as an administrative duty linked to record keeping and accountability. The data

from 2006 suggests that their perspective on what I have termed a 'process versus product' approach impacted on the systems they developed. In a minority, there was evidence of periodic pupil self-reflection on progress using pupil progress files to record attainment, which was also then recorded in teachers' files. However, this practice is not widely evidenced across the schools in the Riverside Partnership. On the other hand, the majority of departments used either a departmental or a school wide database to record and collate interim results of ongoing teacher assessments. However, in some, the complexity of the mathematical manipulation that was built into these systems meant that the dependability of the end of Key Stage attainment levels was flawed.

Practice in the use of target setting is mixed across the schools in the Riverside Partnership. Whilst there is evidence of some meaningful pupil engagement in ongoing target setting, which is linked to pupil self-reflection on progress, the over-simplistic use of sublevels in ongoing target setting within lessons was also evident in the data.

The issues linked to validity and reliability raised in the debate in Chapter Four, regarding the limitations of PE teachers as subject knowledge experts, reinforced the need to develop effective moderation and standardisation processes in order improve the dependability of using ongoing teacher assessment for summative purposes. Whilst the evidence of such developments in the present research is broadly positive, practice varies across the schools in the Riverside Partnership. There are formal and informal approaches to standardisation and moderation across the Riverside Partnership including departmental standardisation and moderation meetings, peer observation of practical performance or observation of video evidence, cross moderation, moderation of video evidence and discussion. In some schools, the Head of Department has an overall moderation and standardisation role. In the data for 2006 only, there was some evidence of co-marking as part of the moderation systems starting to emerge. Whilst most schools recognised the importance in including some or all of these features in their practice, there is evidence that even by 2006, a minority do

not feel these are a necessity. They contend that their length of service as a teacher and the time they have worked with their colleagues in a particular school, is sufficient to assure the quality of the reliability and validity of their assessment practice.

The processes developed for standardisation and moderation often focus on performance, either using videos or observing live performances. This results in an over emphasis on performance, with very little evidence in the data that standardisation and moderation of pupil's abilities to plan and evaluate occurs.

There was some evidence in the present research that the complexity of the construct of PE and how teachers interpret it may affect the dependability of their summative assessment practice, particularly where their own interpretation of the construct was at odds with that defined in the prevailing National Curriculum. This is best evidenced in those schools where it was reported that end of unit attainment levels were commonly decided, based on an observation of a final performance. As it is only possible to observe that which can be seen, summative assessment judgements, using this methodology frequently overemphasise performance. There is no evidence to suggest that this is a conscious effort by the teacher to subvert the assessment process; indeed in some schools it was reported that recorded evidence, (video or digital photos) was shared in the department to help ensure validity and reliability of the judgements made. However, as this focused solely on the final performance, no evidence of assessment or moderation was available regarding the thinking skills required to plan and evaluate. So, even if, in such schools, these grades were systematically collated and mathematically manipulated to reach a final End of Key Stage attainment level, the process of getting there was essentially flawed and the dependability of the assessment was affected.

Conclusions

This longitudinal study, into assessment practice in PE, was undertaken at a particular time, 2000-2006 within a particular policy context, NCPE (2000) and Ofsted (2003b). At the time of the study, the PE teachers in Riverside Partnership were working in schools, where the prevailing performativity and accountability agendas influenced all aspects of education policy and practice. At this time, there was an unprecedented focus on teachers' assessment practice at national level through SNS. This was underpinned by the research of leading theoreticians of the day (Black and Wiliam, 1998a, 1998b; ARG, 1999 – 2010). It is, therefore, unsurprising that what we can see from the data collected for the present study is that PE teachers' practice changed in a number of ways.

The most significant change is in the range of methods used by the teachers to reach dependable judgements in relation to the end of Key Stage 3 attainment levels. Whilst the data suggests that teacher observation continues to be an important part of the PE teachers' assessment practice in 2006, we can see that throughout the study period, PE teachers are increasingly using a wider range of methods and tools, in order to make their judgements. In this performativity climate, with the need to achieve successful outcomes in Ofsted inspections and influenced by the SNS, there is evidence in the present study that the teachers practice moved towards the notion of 'good practice' in assessment in PE as defined by Ofsted (2003b), particularly in relation the range of methods used.

However, one of the consequences of this change in practice, which is relevant to today (2011), has been the change noted in the use of end of Key Stage attainment levels. Within this climate of accountability and performativity and influenced by the prevailing assessment culture in their schools in other subjects, we can see that many PE teachers are using these levels in the way that they were never intended to be used (QCA, 1999). It is possible to see that during this longitudinal study, some PE teachers'

practice was changing, in a way that eventually led to the concerns about teaching to the levels that have been raised by Frapwell (2010).

Though not always mentioned specifically by name, there is evidence of 'Assessment for learning' (Black and Wiliam 1998a) principles being adopted in many schools in Riverside Partnership by 2006. Practice noted includes

- Feedback to inform learning and progress
- Shared criteria for assessment in language pupils understand, including sheets, displays and pupil progress files
- Question and answer to check understanding and inform future learning
- Peer and self-assessment opportunities.

Whilst we can see that some teachers in Riverside Partnership are increasingly using these AfL approaches, it is not possible to assess the extent to which they were being used effectively to develop learner autonomy (Black and Wiliam, 2009) or in a mechanistic way (James, 2006) due to the methodology and timing of the data collection for the research.

We can also see that the complexity of the construct of PE and how teachers interpret, it may affect the dependability of their summative assessment practice, particularly where their own interpretation of the construct is at odds with that defined in the prevailing NCPE (2000). At the time of the study, the prevailing conceptualisation of PE as represented by NCPE (2000) was an educative rather than a Sport construct. This is of interest today, in that the Sport activities have been completely driven out of NCPE (2008), which focuses on the development of cognitive skills (key concepts and key processes) through a range of practical contexts.

Finally, it can be concluded that PE teachers in the schools in the Riverside Partnership use a general 'best-fit' approach to determine pupils' summative attainment levels at the end of Key Stage 3, in a similar way to the teachers from other subjects (Clarke and Gipps, 1998). However, there is some

inconsistency across the partnership in the ways in which teachers are interpreting what a child needs to do to evidence their achievement of a particular level, which has implications for the quality of training offered to the student teachers in Riverside.

Having presented the main conclusions for the thesis, this section offers some implications for policy and practice in Riverside and future research.

Implications for policy and practice in Riverside

The university and the schools in Riverside Partnership need to:

1. Ensure that all trainees develop a good theoretical understanding of teacher assessment issues in order to develop dependable assessment in PE for a variety of purposes by the end of their PGCE course.

2. Ensure that all trainees experience and develop a wide range of assessment methods, as part of the training through their PGCE course.

3. Consider ways of developing the PE trainees understanding of the complexity of the construct of PE, and how their values can impact on their assessment practice.

4. Provide opportunities for the trainees to consistently observe 'good practice' in teacher assessment for a variety of purposes whilst on school placement, including how teachers make 'best-fit' judgements at Key Stage 3.

5. Continue to work with other teacher education providers in PE at a regional and national level to share and disseminate how to ensure dependable assessment in PE at Key Stage 3.

Suggestions for related future research

Having completed this study into PE teachers' assessment practice, there are a number of areas of interest that I would like to explore. Two are linked to assessment practice in PE; the final one arises out of my growing interest in the policy context, in which this study took place and its power to change teachers practice.

1. Evaluation of the impact on the practice of PE teachers of the new 'Assessing Pupil Progress' (QCA, 2009) once it has been fully developed and implemented in schools.

2. The relationship between teachers' constructs of PE and the dependability of their assessment practice.

3. An investigation into how national education political agendas drive changes into teachers' practice

Chapter Six: Postscript to a Thesis

This study has been an important part of my life for over a decade. As a result, one of the key difficulties I have experienced in these final months has been to know when to stop! This last decade, since the study was conceived in 1998, has seen many developments in assessment practice at national level, and looking back through the data collected for the present study, serves to remind me how practice in this area has evolved.

The present research focused on the change in assessment practice of PE teachers between 2000 and 2006 in Riverside Partnership. However, although 2006 was a cut off point for the data collected for the present study, it would be inappropriate to suggest that the pace of developments in assessment at a national level has slowed down. Indeed, since 2006 the focus on assessment within the SNS has sharpened. The main research findings of the present study provide evidence of change in assessment practice in PE in many schools across the partnership, with teachers moving towards the notion of 'good practice' in assessment in PE promoted by Ofsted (2003b). There is clear evidence that, whilst teacher observation continues to be an important part of their overall assessment strategy, the PE teachers, in Riverside Partnership now use a wider range of methods to inform their assessment judgements. The programme of CPD, which supported the implementation of the SNS, coupled with the Ofsted inspection regime has influenced these changes in the PE teachers' assessment practice.

At a national level, there is evidence to suggest that these developments in assessment in PE are continuing, although, according to Ofsted, there are still areas requiring improvement. In 2009, following the most recent review of practice in a sample of primary and secondary schools in PE from 2005 to 2008 Ofsted (2009, p.5) concluded that:

> *The better schools visited, assessed, recorded and tracked pupils' progress systematically. However, because there is no*

common assessment strategy nationally, inconsistencies remained in judging pupils standards and achievements accurately.

This conclusion is of interest to me, for whilst this 3-year evaluation of PE in 99 primary schools and 84 secondary schools was not completed until 2008, it commenced during the time of the present study. Of particular interest is the key concern, raised by Ofsted, of a lack of a 'common assessment strategy nationally'. In 2008, QCA in partnership with the SNS began to develop a national approach to assessment known as Assessing Pupils' Progress project (APP):

Assessing pupils' progress (APP) is a national approach to assessment that equips teachers to make judgements on pupils' progress [...] APP helps teachers to fine-tune their understanding of pupils needs and tailor their planning and teaching accordingly, by enabling them to...make reliable judgements related to national standards drawing on a wide range of evidence (QCA, 2009, p.1).

This project, like the SNS, cites the work of the ARG (1999 to 2010), as its theoretical underpinning. However, writing in 2009, Marshall, James and the ARG suggest that this government-developed version of assessment for learning:

...shares little of the "spirit" of the definition and principles from the Assessment Reform Group, although the documentation quotes them. Indeed, Assessing Pupils' Progress, the in-class assessment system that is a part of the government's version of assessment for learning in England, is more to do with specifying frequent summative assessment than formative assessment (p.28).

The potential impact of this national project, which is being rolled out through the core and foundation subjects of the National Curriculum, is of interest to me both professionally and for my future research activity. It is also of interest, as it appears to be a further example of the work of leading theoreticians of the day being mediated by the policy makers to drive wholesale changes in teachers' classroom practice in a particular way. Whilst the legitimacy of political involvement in education policy is not in doubt, as Mansell, James and the ARG (2009, p.28) observe:

> *While no one would contest the right of elected politicians to determine overall assessment policy, their involvement in specifying technical details of assessment models and procedures raises questions over whether they, and some of their advisers, are sufficiently qualified to do so at such a detailed level.*

Since this study was completed the NCPE (2000) has been revised, and the new version, implemented in 2008 is based on an educational, rather than a sporting construct of PE. Indeed, the sport activities have now completely disappeared from the documentation, being replaced by key concepts and key processes that must be taught through practical contexts. Thus, whilst pupils are required to "outwit opponents" or undertake "movement replication", there is no mention of any specific activities, games or sports by name. The impact of this view of knowledge in the revised curriculum (NCPE, 2008) on teachers' assessment practice remains to be seen. Whilst this would be of interest to me for future research, the Conservative led coalition government, which took over from New Labour in May 2010, is already proposing that a much-reduced National Curriculum will be implemented from 2013, in which PE may not even be included.

In this postscript, I have reflected on my experience of undertaking this study and what I have learned from taking part in this Doctoral programme. I have definitely come to appreciate just how difficult it is to do research at this level with a full time job and young children. My daughters were 2 and

5 respectively when I started, 15 and 17 by the time I finished. On the one hand, my work and family commitments have definitely increased the time it has taken me to complete this study. On the other hand, this has meant that data is available for a seven-year period. As a result, it has been possible to look at how the assessment practice of the PE teachers in Riverside Partnership has changed over this period, against a background of so many developments in assessment practice at national level, within the policy context of the NCPE (2000) and the Ofsted (2003b) notion of 'good practice' in assessment in PE. I feel I have matured as a researcher and my commitment to mixed methodologies has strengthened. The insights offered through analysing the commentaries have been fascinating. I am also very conscious of my own role in the research process, in terms of my potential to unwittingly affect the research outcomes, Helliwell's 'insider-outsider perspective' (2006, p.488).

On a professional level, the lessons learned though my engagement with this Doctoral thesis continue to inform my own professional practice in many ways. These include my subject knowledge for my lectures and seminars on assessment as part of the Post Graduate Certificate in Education PE and my capability as a research supervisor, for the undergraduate courses, on which I teach.

Finally, as with any research, there is much that with hindsight I might change, for example the scope of the literature review, the research design or the way the data was analysed. Whilst these changes cannot be made for the present study, the lessons learned will be used to shape and inform my future research practice.

References

Alderson, G. J. K. and Crutchley, D. H. (1990) 'Physical Education in the National Curriculum', in Armstrong, N. (eds.) (1990) *New Directions in PE,* vol. 1. Human Kinetics

Armstrong, N. (1992) *New Directions in Physical Education*, vol. 2 Leeds, Human Kinetics

Armstrong, N. and Sparkes, A. (1991) *Issues in Physical Education.* London, Cassell

Assessment Reform Group (1999) *Assessment for Learning Beyond the Black Box,* University of Cambridge, Assessment Reform Group.

Assessment Reform Group (2002) *Assessment for learning: 10 principles,* University of Cambridge, Assessment Reform Group

Assessment Reform Group (2003) 'The role of teachers in the assessment of learning' *Pamphlet from Assessment Systems for the Future Project* Nuffield Foundation, Assessment Reform Group.

Atkinson, P., Delamont, S. and Hammersley, M. (1988) 'Qualitative research traditions' in Hammersley, M. (ed.) (1993) *Educational Research: Current Issues*, vol. 1, London, Paul Chapman Publishing/The Open University.

Bailey, R. P. (2005) 'Evaluating the Relationship between Physical Education, Sport and Social Inclusion', *Educational Review*, vol. 57 no.1, pp. 71–90.

Bailey, R. P. and Dismore, H. (2004) 'Sport in Education (SpinEd) – the role of physical education and sport in education', Project Report to the 4th

International Conference of Ministers and Senior Officials Responsible for Physical Education and Sport, Athens, Greece: MINEPS IV.

Bailey, R. P., Dismore, H. and Morley, D. (2009) 'Talent development in physical education: a national survey of practices in England', *Physical Education and Sport Pedagogy*, vol.14, no.1, pp. 59–72.

Bailey, R. and Morley, D. (2008) Physical Education Quality Standards for Talent Development

Ball, S. J. (1990) 'Self-doubt and soft data: social and technical trajectories in ethnographic fieldwork', in Hammersley, M. (ed.) (1993) *Educational Research: current issues*, vol.1, London, Paul Chapman Publishing/The Open University.

Ball, S. J. (2000) 'Perfomativities and fabrications in the education economy: Towards the performative society', *Australian Educational Researcher,* vol. 27 no.2, pp.1-24.

Ball, S. J. (2003) 'The teacher's soul and the terrors of performativity', *Journal of Education Policy*, vol.18, no. 2, pp. 215-228.

Ball, S. J. and Gewirtz, S. (2000) 'Schools, cultures and values: the impact of the 1988 and 1993 Education Acts', London, ESRC.

Ball, S. J. (1994) *Education Reform: A critical and post-structuralist approach*, Buckingham, Open University Press,

Bell, J. (2005, 4th edn) *Doing Your Research Project: A Guide for First-Time Researchers in Education, Health and Social Science* London, Open University Press.

Bell, J., Bush, T., Fox, A., Goodey, J. and Goulding, S. (eds) (1984) *Conducting Small-Scale Investigations in Educational Management,* London, Harper and Row.

Benett, Y. (1993) 'The Validity and Reliability of Assessments and Self Assessments of Work Based Learning', in Murphy, P. (ed) (1999) *Learners, Learning and Assessment,* Paul Chapman Publishing / The Open University.

Bird, M. and Hammersley, M. (1995) E835: *Offprints Reader.*

Black, P. (1995) 'Can Teachers Use Assessment To Improve Learning?' *British Journal of Curriculum and Assessment,* vol. 5, no. 2, pp.7-11.

Black, P. (1996) 'Meanings and consequences: a basis for distinguishing formative and summative functions of assessment?' *British Educational Research Journal,* vol. 22, no.5, pp. 537-546.

Black, P. (1998) 'Assessment, Learning Theories and Testing Systems', in Murphy, P. (eds) (1999) *Learners, Learning and Assessment,* Paul Chapman Publishing/The Open University.

Black, P. (2005) 'Formative assessment: views through different lenses' In *The Curriculum Journal,* vol. 16, no. 2 pp. 133-135 Routledge.

Black, P. and Wiliam, D. (1998a) 'Assessment and classroom learning'. *Assessment in Education, vol.*5, no.1, pp. 7 – 74.

Black, P. and Wiliam, D. (1998b) *Inside the Black Box: Raising standards through classroom assessment,* London, Kings College

Black, P. and Wiliam, D. (1998c) *Beyond the Black Box,* University of Cambridge, School of Education.

Black, P. and Wiliam, D. (2008) Developing the theory of formative assessment, *Educational Assessment, Evaluation and Accountability* February 2009, vol. 21, no 1, pp. 5-31.

Black, P., Harrison, C., Lee, C., Marshall, B. and Wiliam, D. (2003) *Assessment for Learning: Putting it into practice*, Maidenhead, Open University Press.

Black, P., Harrison, C., Hodgen, J., Marshall, M. and Serret, N. (2010) 'Validity in teachers' summative assessments', *Assessment in Education* vol. 17, no.2, pp. 215-232.

Black, P., McCormick, R., James, M. and Pedder, D. (2006) 'Learning How to Learn and Assessment for Learning: a theoretical inquiry', *Research Papers in Education,* vol. 21, no. 2, pp. 119 – 132.

Borg, W.R. and Gall, M. D. (1989) *Educational Research: An introduction* (5th edn) London, Longman.

Brannen, J. (1992) 'Combining qualitative and quantitative approaches: an overview', in Brannen, J. (ed) *Mixing Methods: Qualitative and Quantitative Research.* Aldershot, Avebury.

British Educational Research Association (2004) *Revised guidelines for Educational Research* [pdf] available at: http://www.bera.ac.uk/files/ guidelines/ethica1.pdf [accessed 18 June 2010].

Broadfoot, P. (1979). *Assessment, Schools and Society Contemporary Sociology of the School*, London, Methuen Publishing Ltd.

Broadfoot, P. (1996) *Education, Assessment, and Society,* Milton Keynes, Open University Press.

Broadfoot, P. (2000a) 'Preface', in Filer, A. (eds), *Assessment: Social practice and social product* (pp. ix-xii). New York: Routledge Falmer

Broadfoot, P. (2000b) 'Empowerment or performativity? Assessment policy in the late twentieth century', in Phillips, R. and Furlong, J. (eds), *Education, Reform and the State: Twenty-five years of politics policy and practice*. London, Routledge Falmer.

Broadfoot, P. (2007) *An Introduction* to *Assessment*, Continuum International Publishing Group, New York.

Bryman, A. (1988) *Quantity and Quality in social Research,* London, Unwin Hyman.

Bryman, A. (1992) Quantitative and qualitative research: further reflections on their integration, in Brannen, J. (ed.) *Mixing Methods: Qualitative and Quantitative Research*, Aldershot, Avebury.

Capel, S. (1997) *Learning to Teach Physical Education in the Secondary School,* London, Routledge Falmer.

Capel, S. (2000) 'Physical Education and Sport' in Capel, S. and Piotrowski, S. (ed) *Issues in Physical Education,* London, Routledge.

Capel, S., Leask, M. and Turner, T. (eds.) (2009) 'Introduction' in *Learning to teach in the secondary school: A companion to school experience* (5th edn), London, Routledge.

Carr, M., McGee, C., Jones, A., McKinley, E., Bell, B., Barr, H. and Simpson, T. (2000) *Strategic research initiatives: the effects of curricula and assessment on pedagogical approaches and on educational outcomes* Ministry of Education, Wellington, New Zealand.

Carroll, B. (1991) 'Assessment in the National Curriculum: What the teacher has to do', *British Journal of Physical Education,* vol. 22, no. 2, pp. 8-10.

Carroll, B. (1994) *Assessment in Physical Education,* London, The Falmer Press.

Casbon, C. and Spackman, L. (2005) 'Assessment for Learning in Physical Education' – BAALPE.

Clarke, S. and Gipps, C. (1997) *Evaluation of Key Stage 1 Statutory Assessment (England),* SCAA, London.

Clarke, S. and Gipps, C. (1998) 'The Role of Teachers in Teacher Assessment in England 1996-1998', *Evaluation and Research in Education* vol. 14, no. 1, pp. 38–52.

Cohen, L. and Manion, L. (1980) *Research Methods in Education,* London, Croom Helm.

Cohen, L., Manion, L. and Morrison, K. (1996) *A Guide to Teaching Practice*, New York, Routledge.

Coladarci, T. (1986) 'Accuracy of teachers' judgements of students responses to standardised test items', *Journal of Educational Psychology* vol. 78, pp, 141-146.

Datta, L. (1994) 'Paradigm wars: A basis for peaceful coexistence and beyond', in Reichardt, C.S. and Rallis, S.F. (eds) *The qualitative-quantitative debate: New perspectives*, San Francisco, Jossey-Bass.

DCMS (2000) *A Sporting Future for All.* London, DCMS

Denscombe, M. (1998) *The Good Research Guide,* Buckingham, Open University Press.

Department for Education and Employment (DfEE) (1997) *Excellence in Schools* (White Paper), London, HMSO.

Department for Education and Employment (DfEE) (2000) *Physical Education in the National Curriculum,* London, HMSO.

Department of Education and Science (1988) *Task Group on Assessment and Testing A Report,* London, HMSO.

Department of Education and Science / QCA (2001) Key Stage 3 National Strategy Key messages about assessment for learning [online], http://www.efl.org/custom/resources_ftp/client_ftp/teacher/mfl/ks3mfl/key_messages/Assessment%20for%20Learning.pdf (accessed 11 August 2010).

DES and Welsh office (1991) *Physical Education in the National Curriculum,* London, HMSO.

DfEE and WELSH OFFICE (1995) *Physical Education in the National Curriculum,* London, HMSO.

Dickenson, B. and Almond, L. (1993) 'Assessment in Physical education: A practical model for implementing National Curriculum Assessment requirements' *British Journal of Physical Education,* vol. 24, no.4, pp 22-26.

Drewett, P. (1991) 'Assessment developments and challenges in Physical Education', in Armstrong, N. and Sparkes, A. (eds) *Issues in Physical Education,* London, Cassell Education.

Easterby-Smith, M., Thorpe, R. and Lowe, A. *(2002) Management Research: An Introduction, 2nd Edition,* London, Sage Publications.

Eisner, E. (1992) 'Objectivity in educational research', in Hammersley M (eds.) (1993) *Educational Research: current issues*, vol. 1, London, Paul Chapman Publishing / The Open University.

Ercikan, K., and Roth, W. M. (2006) 'What good is polarizing research into qualitative and quantitative?' *Educational Researcher*, vol. 35, no. 5, pp.14-23.

Evans, J. and Penney, D. (1995) 'The politics of pedagogy: making a National Curriculum Physical Education', *Journal of Educational Policy,* vol. 10, no. 1, pp. 27-44.

Faulkner, G. and Sparkes, A. (1999), 'Exercise as therapy for schizophrenia: An ethnographic study'. *Journal of Sport and Exercise Psychology,* vol. 21, pp.52-69.

Freebody, P. (2003) *Qualitative Research in Education,* London, Sage.

Filer, A. and Pollard, A. (2000) *The social world of pupil assessment: processes and context of primary schooling*', New York, Continuum.

Foucault, M. (1974) *Power/knowledge: Selected interviews and other writings*, New York, Pantheon Books.

Frapwell, A. (2010) Assessment in physical education – Fit *for* purpose? We've got an APP *for* that! *Physical Education Matters*, vol. 5, no. 3, pp. 13-18.

Gardner, H. (1992) 'Assessment in context', in Murphy, P. (ed.) (1999) *Learners, Learning and Assessment,* Paul Chapman Publishing/The Open University.

Gardner, J. (ed.) (2006) *Assessment and Learning.* London, Sage.

Goetz, J. P. and LeCompte, M. D. (1984) *Ethnography and qualitative design in educational research,* Orlando, Academic Press.

Gipps, C. (1990) *Assessment: A Teachers' Guide to the Issues,* Hodder and Stoughton.

Gipps, C. (1999) 'Socio-cultural aspects of assessment', *Review of Research in Education,* vol. 24, pp. 355-392.

Gipps, C. and Clarke, S. (1996) *Monitoring consistency in Teacher Assessment and the Impact of SCAA's Guidance Materials at Key Stages 1,2 and 3,* SCAA, London.

Gipps, C., Clarke, S. and McCallum, B. (1998) *The Role of Teachers in National Assessment in England.* Paper presented at the annual meeting of the American Research Association, San Diego CA .

Gold, R. L. (1958) 'Roles in sociological field observations', *Social Forces,* vol. 36, no. 3, pp. 217-233.

Gordon, E. W. (2008) 'The transformation of key beliefs that have guided a century of assessment', in Dwyer C, A. (eds.), *The future of assessment,* pp. 3-6, New York, Taylor and Francis Group.

Gough, D. (2006) *User led research synthesis: a participative approach to driving research agendas.* Presented at the Sixth International Campbell Colloquium, Los Angeles, 22–24 February.

Green, K., Smith, A. and Roberts, K. (2005) 'Young people and lifelong participation in sport and physical activity: a sociological perspective on contemporary physical education programmes in England and Wales', *Leisure Studies,* vol. 24, no.1, pp. 27–43.

Guba, E. G. (2005) 'Paradigmatic controversies, contradictions, and emerging confluences', in Denzin, N. K. and Lincoln, Y. S. (eds.) *Handbook of Qualitative Research,* Third edition, Thousand Oaks CA, Sage.

Guba, E. G. and Lincoln, Y. S. (1989) *Fourth Generation Evaluation.* Newbury Park, CA: Sage.

Guba, E. G. and Lincoln, Y. S. (1995) 'Competing paradigms in qualitative research', in Denzin, N. K. and Lincoln, Y. S. (eds.) *Handbook of Qualitative Research* , Thousand Oaks CA, Sage.

Hammersley, M. (1984) 'The researcher exposed: a natural history' in Burgess, R. G. (ed.) *The Research Process in Educational Settings: Ten case studies,* Lewes, Falmer Press.

Hammersley, M. (1992) 'The Paradigm Wars: Reports From the Front'. *British Journal of Sociology of Education*, vol.13, no.1, pp.131–143.

Hammersley, M. (ed.) (1993) *Educational Research: current issues*, vol. 1, London, Paul Chapman Publishing/The Open University.

Hammersley, M. (1995) 'Opening Up the Quantitative-qualitative Divide'. *Education Section Review*, vol. 19, no.1, pp. 2-15.

Hammersley, M. and Atkinson, P. (2005, 2nd edn) *Ethnography: Principles in Practice*, London, Routledge.

Hammersley, M. and Gomm, R. (2000) ' Introduction' in Gomm, R., Hammersley, M. and Foster, P. (eds) *Case Study Method*, Thousand Oaks, CA, Sage.

Harlen, W. (2004a) 'A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes', (EPPI-

Centre Review) in *Research Evidence in Education Library,* issue 3. London, EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2004b) 'A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes' (EPPI-Centre Review), in *Research Evidence in Education Library, issue* 4,London, EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2005a) 'Teachers' summative practices and assessment for learning - tensions and synergies'. *The Curriculum Journal,* vol. 16 no. 2 pp. 133-135, Routledge.

Harlen, W. (2005b) 'Trusting teachers' judgements: research evidence of the reliability and validity of teachers' assessment used for summative purposes', *Research Papers in Education*, vol. 20, no. 3, pp. 245 – 270.

Harlen, W. (2009) 'Improving assessment of learning and for learning'. *Education 3-13*, vol. 37, no. 3, pp. 247-257.

Harlen, W. and Deakin Crick, R. (2002) 'A systematic review of the impact of summative assessment and tests on students' motivation for learning' in *Research Evidence in Education Library*, London, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W. and James, M. J. (1997) 'Assessment and learning: differences and relationships between formative and summative assessment', *Assessment in Education*, vol. 4, no.3, pp. 365–80.

Harris, D. and Bell, C. (1990) *Evaluating and Assessing for Learning* London, Kogan Page Ltd.

Harrison, J., Blakemore, C., Buck, M. and Pellett, T. (1996) *Instructional Strategies for Secondary School Physical Education,* London, Brown and Benchmark.

Hellawell, D. (2006) 'Inside –out: analysis of the insider-outsider concept as a heuristic device to develop reflexivity in students doing qualitative research', *Teaching in Higher Education,* vol. 11, no. 4, pp. 483 – 494.

Houlihan, B. (2000) 'Sporting excellence, schools and sports development: the politics of crowded policy spaces', *European Physical Education Review*, vol. 6, no. 2, pp. 171 - 193.

James, M. (2006) "Assessment, teaching and theories of learning". in Gardner, J. (ed.) *Assessment and Learning*, London, Sage.

Kerlinger, F. H. (1986, 3rd edn) *Foundations of behaviour research*, New York, Holt Reinhart and Winston

Kirk, D. (2010) 'Defining physical education. Nature, purpose and future/s', *Physical Education Matters*, vol. 5, no. 3, pp. 13-18.

Kirk, D. (2005) 'Physical education, youth sport and lifelong participation: the importance of early learning experiences', *European Physical Education Review*, vol. 11 no. 3, pp. 239–55.

Kirk, D. and Gorely, T. (2000) 'Challenging thinking about the relationship between school physical education and sport performance', *European Physical Education Review*, vol. 6, no. 2, pp. 119–34.

Kirk, D. and Tinning, R. (1990) Physical Education, Curriculum and Culture: Critical issues in the Contemporary Crisis London: Falmer

Knudson, D. and Morrison, C. (2002, 2nd ed) *Qualitative Analysis of Human Movement.* Champaign, IL, Human Kinetics.

Kolb, D. A. (1984) *Experiential Learning: Experience on the Source of Learning and Development,* Englewood Cliffs: Prentice Hall.

Kvale, S. (1995) 'The social construction of validity' *Qualitative Inquiry,* vol.1, no.1, pp. 19-40.

Lee, M. J. (2004) 'Values in physical education and sport: a conflict of interests' *The British Journal of Teaching Physical Education,* vol. 35, no.1, pp. 6-8.

Lincoln, Y. S. and Guba, E. G. (1985) *Naturalistic inquiry.* Beverly Hills: Sage.

Lockwood, A. (2000) 'Breadth, and Balance in the Physical Education Curriculum', in Piotrowski S and Capel S (eds) (2000) *Issues in Physical Education,* Routledge, London.

McChonachie-Smith, J. (1991) 'Assessment of progression in National Curriculum Physical Education', *British Journal of Physical Education,* vol. 22, no. 2, pp. 11-15.

MacDonald and Brooker (1997) 'Assessment Issues in a Performance Based Subject- A Case Study of PE', in Murphy P (eds) (1999) *Learners, Learning and Assessment,* Paul Chapman, Publishing/The Open University.

Mansell, W., James, M. and the Assessment Reform Group (2009). *Assessment in Schools. Fit for purpose? A commentary by the Teaching and Learning Research Programme.* (Booklet.) London, ESRC TLRP, Institute of Education London.

Marshall, B. and Drummond M. J. (2006) 'How teachers engage with Assessment for Learning: lessons from the classroom', *Research Papers in Education,* vol. 21, no. 2, pp. 133–149.

Marshall, C. and Rossman, G. B. (2006, 4th edn.) *Designing Qualitative Research*, Thousands Oaks: Sage Publication.

Mawer, M. (1995) *The Effective Teaching of Physical Education*, Longman.

Maxwell, G. and Gipps, C. (1996) 'Teacher Assessments of Performance Standards: A Cross-National Study of Teacher Judgements of Student Achievement in the Context of National Assessment Schemes'. Application for funding to the ARC, Interdisciplinary and International Research.

May, T. (1997) *Social research*, Buckingham, Oxford University Press.

Merton, R. (1972) 'Insiders and outsiders: a chapter in the sociology of knowledge', *American Journal of Sociology*, vol. 78, pp. 9-47.

Morley, D. (2008) 'Viewing physical education through the lens of talent development', unpublished doctoral dissertation, Leeds Metropolitan University.

Morley, D. and Bailey, R. P. (2006) *Meeting the Needs of Very Able Pupils in Physical Education and Sport*, London, David Fulton.

Morley, D., Bailey, R.P. and Cobley, S. (2006) 'Pupils perceptions of talent in physical education', a report of the *British Educational Research Association,* Warwick University.

Morrison, K. (1996) 'International developments in assessment', in Craft, A. (eds.) *Primary education: Assessing and planning learning,* New York, Routledge.

Murdoch, E. (1990) 'Physical Education and Sport: the interface', in Armstrong, N. (eds) (1996) *New Directions in Physical Education* vol. 1, pp. 63-79, Champaign, Ill, Human Kinetics.

Nelligan, C. (1995) 'Assessment in Physical Education', *The Bulletin of Physical Education,* vol. 31, no. 3, pp.10-16.

Newton, P. (2010) 'Educational Assessment – Concepts and Issues: The Multiple Purposes of Assessment', in Baker, E., McGraw, B. and Peterson, P. (eds.) *International Encyclopedia of Education, Third Edition.* Oxford.

Office for Standards in Education (1993) *Assessment recording and reporting: Third year 1991 –1992,* Ofsted Publications, London, HMSO.

Office for Standards in Education (1995) *Physical Education: A Review of Inspection Findings 1993/94,* Ofsted Publications, London, HMSO.

Office for Standards in Education (1998) *Review of subject inspections in ITT (1996 - 1998),* Ofsted Publications, London, HMSO.

Office for Standards in Education (2002) *Physical education in secondary schools subject reports series 2001/02*, Ofsted Publications, London, HMSO.

Office for Standards in Education (2003a) *Good assessment in secondary schools*, Ofsted Publications, London, HMSO.

Office for Standards in Education (2003b) *Good assessment practice in physical education*, Ofsted Publications, London, HMSO.

Office for Standards in Education (2005) *The Secondary National Strategy An evaluation of the fifth year.* Ofsted Publications, London, HMSO.

Office for Standards in Education, Children's Services and Skills (2009) *Physical education in schools 2005/08 Working towards 2012 and beyond.* Ofsted Publications, London, HMSO.

Patton, M. Q. (1980) *Qualitative Evaluation Methods*, New York, Sage Publications.

Patton, M.Q. (1990, 2nd edn) *Qualitative evaluation and research methods,* Newbury Park, California. Sage.

Penney, D. (2000) 'Physical education, sporting excellence and educational excellence', *European Physical Education Review,* vol.6, p.2, pp. 135–150.

Piotrowski, S. and Capel, S. (2000) 'Formal and informal modes of assessment' in physical education', in Capel, S. and Piotrowski, S. (eds) (2000) *Issues in Physical Education,* Routledge, London.

Qualifications and Curriculum Authority (QCA) (2001). *Assessment and Reporting Arrangements 2001: Key Stage 3*. London. HMSO.

Qualifications and Curriculum Development Agency (2009) The National Strategies Assessing Pupils' Progress:[online], http://nationalstrategies.standards.dcsf.gov.uk/node/259613 (accessed 8 June 2010).

Reichardt, C. S. and Rallis, S. F. (1994) 'Qualitative and quantitative inquiries are not incompatible: A call for a new partnership', in Reichardt, C.S. and Rallis, S.F. (eds) *The qualitative-quantitative debate: New perspectives*, San Francisco, Jossey-Bass.

Reynolds, D. (2002) 'Developing differently: educational policy in England, Wales, Scotland and Northern Ireland', in. Adams, J. and Robinson, P. (eds) *Devolution in Practice*. IPPR, London.

Roy, W. (1983) *Teaching under Attack*, London, Croom Helm.

Sadler, D. R. (1989) 'Formative Assessment and the design of instructional systems', *Instructional Science,* no. 18, pp. 119-144.

Satterly, D, (1981) *Assessment in School*, London, Muslim Education Trust.

Schofield, J. (1993) 'Increasing the generalizability of qualitative research' in M Hammersley (ed) *Social research: Philosophy, Politics and Practice,* London, Open University and Sage.

School Curriculum and Assessment Authority (1997) *Physical education at Key Stages 3 and 4 Assessment Recording and Reporting: Guidance for Teachers*, London, SCAA Publications.

Shulman, L. S. (1987) 'Knowledge and teaching: foundations of the new reform', *Harvard Educational Review*, no. 57, pp. 1–22.

Simon, A., Sohal, A. and Brown, A. (1996) 'Generative and case study research in quality management, Part 1: Theoretical considerations', *International Journal of Quality and Reliability*, vol. 13, no. 1, pp. 32 –42.

Stake, R. E. (1978) 'The case study method in social inquiry', *Educational Researcher,* vol 7, no.2, pp 5-8.

Stake, R. E. (1995) *The Art of Case Study Research,* Thousand Oaks, CA Sage.

Spenceley, P. (2009) '10 years of AFL – A personal view' [online] Curriculum *Management Update*, November 2009, Issue 100 http://www.curriculum-management-update.com/article/assessing-impact-assessment-learning-10-years# (accessed 2 December 2010).

Stake, R. E. (1998) 'Case Studies', in Denzin N.K and Lincoln Y.S (eds) *Strategies of Qualitative Inquiry*, Thousand Oaks, CA: Sage.

Stobart, G. (2001) The Validity of National Curriculum Assessment, *British Journal of Educational Studies,* vol. 49, no. 1, pp. 26-39.

Stobart, G. (2008) *Testing Times: the uses and abuses of assessment,* Abingdon, Routledge.

Tanner, H. and Jones, S. (1994). 'Using peer and self-assessment to develop modelling skills with students aged 11 to 16: a socio-constructive view', *Educational Studies in Mathematics*, vol. 27, no.4, pp. 413-431.

Tashakkori, A. and Teddlie, C. (1998) *Mixed Methodology: Combining Qualitative and Quantitative Approaches,* Thousand Oaks, CA, Sage.

TGAT (1988) *A Report,* London, DES/WO.

The Open University (1990) E819 *Curriculum, Learning and Assessment,* Milton Keynes, Hobbs Hampshire The Open University.

The Open University (1996) E835, *Educational Research in Action, Study guide Section* 6, *Qualitative Research,* Hobbs: Milton Keynes, Hobbs Hampshire The Open University.

Whitehead, M. E. (2000) 'Aims as an issue in physical education' in Piotrowski, S. and Capel, S. (eds) (2000) *Issues in Physical Education,* Routledge, London.

Whitehead, M.E. (2007) 'Physical Literacy: Philosophical considerations in relation to the development of self, universality and prepositional knowledge', *Sport, Ethics and Philosophy,* vol.1, no. 3, pp. 281-298.

Whitehead, M. E. and Murdoch, E. 'Physical Literacy and Physical Education: Conceptual Mapping', *Physical Education Matters,* Summer 2006.

Wiggins, G. (1989) Teaching to the (Authentic) Test. *Educational leadership,* vol.46, no. 7, pp. 41-47.

Wiggins, G. P. (1998) *Educative assessment: Designing assessments to inform and improve student performance*, San Francisco, Jossey-Bass Publishers.

Wiggins, G. P. and McTighe, J. (2006) 'Examining the teaching life', *Educational Leadership,* vol. 63, pp. 26-29.

Williams, A. (1997) *National Curriculum Gymnastics,* London, Hodder and Stoughton.

Winter, G. (2000) 'A comparative discussion of the notion of validity in qualitative and quantitative research', *The Qualitative Report*, vol. 4 no.3 and 4 [online]http://www.nova.edu/ssss/QR/QR4-3/winter.html (accessed 25 June 2010).

Wragg, E.C. (1994) *An introduction to classroom observation,* London, Routledge.

Yin, R. K. (2003) *Case Study Research: Design and Methods*, 3rd edition, Thousand Oaks, California, Sage.

Youngman, M. B. (1982) 'Designing Questionnaires,' in Bell, J., Bush, T., Fox, A., Goodey, J. and Goulding, S. (eds) (1984) *Conducting Small-Scale Investigations in Educational Management*, London, Harper and Row.

Appendices

Appendix One

The evolution of an Ed D thesis between1998 – 2006:
An account of the reasons that lead up to the refocusing of the present research in September 2004

Ofsted (1998) in its report of Initial Teacher Training subject inspections (1996 - 1998) underpinned the widely held view, within the profession that assessment within Physical Education was problematic. This, it suggested was directly linked to the lack of good models of assessment practice within many Physical Education departments in schools. Previously, as a Physical Education teacher and now as a practitioner in initial teacher training in Physical Education, the reasons as to why assessment in Physical Education is problematic have long been of interest to me.

The present study has undergone significant transformation from its original inception. In 1998, when first deciding to investigate the validity of end of Key Stage 3 reports to parents in Physical Education, based on teacher observation, the practice and culture for assessment in this subject at Key Stage 3 was very different from that which exists today. It was commonly accepted that summative assessment in Physical Education could rely on teacher observation, with little or no consideration of the validity of the methodology of such practice. Given a personal dissatisfaction with summative assessment practice in Physical Education at Key Stage 3, the researcher was interested in investigating the validity and reliability of end of Key Stage 3 summative Physical Education reports to parents, based on such teacher observation and the purpose at that time was to prove the hypothesis that;

> *Teacher observation of pupil achievement is subjective and unreliable. Consequently, summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for Physical Education based on solely on an assessment strategy of teacher observation are invalid*

The (1995) National Curriculum for Physical Education (NCPE) was revised in 1999, with significant changes made to the assessment

requirements, criteria and practice for implementation in September 2000. These changes necessitated a different approach to assessment practice in Physical Education, than that which was required under the previous version of the NCPE, (1995).

Logically, such external changes have impacted significantly on all aspects of the current research. The principal decision to change from a positivist hypothesis approach to an interpretive investigation of a primary research question is a major shift for the present study and has resulted in significant changes to the research design including the methodological approaches adopted

Hypothesis approach

> *Hypothesis: Teacher observation of pupil achievement is subjective and unreliable. Consequently, summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for Physical Education based on solely on an assessment strategy of teacher observation are invalid*

.

The main questions that were to be addressed in this research were

1. To what extent is teacher observation subjective?
2. To what extent is teacher observation unreliable?
3. To what extent are summative reports of pupil progress at the end of Key Stage 3 in the National Curriculum for Physical Education (NCPE2000) based on solely on an assessment strategy of teacher observation valid?

Focus of the study: Teacher observation and Teachers' Professional Judgement

The primary issue that concerned me was the subjective nature of such judgements to what extent could such a subjective method be used to gather valid assessment information?

Although, at the time, there was evidence; for example, Mawer (1995) Carroll (1994) Cohen et al. (1996) to suggest that teacher observation was a

useful assessment method in Physical Education, in that its non-intrusive nature ensures that the pupils' performance observed reflects their true ability and that assessment judgements based on such observations are therefore reliable and valid, the primary issue that concerned me was the subjective nature of such judgements. For example the decision about who and what to observe is left to the teacher, and is often made subconsciously, therefore this raises the question is the performance observed by the teacher at the time of observation an example of the pupil's best or worst work? To what extent could such a subjective method be used to gather valid assessment information?

Case study Strategy

In order to test this hypothesis, the reliability and validity of assessment strategies based solely teacher observation, were to be examined in the context of a small number of secondary schools' Physical Education departments. The participants were to be two individuals who were keen to promote teacher observation as totally reliable and valid as a means of reaching summative judgements, with no other method for assessment considered.

Key Changes which led to refocusing in 2004

NCPE revised 1999 for 2000 implementation. Significant changes to the curriculum and even more significant changes to the attainment target and requirements for grading

1995 version of National Curriculum for Physical Education

Planning performing and evaluating with emphasis being on performance.

Assessment made on broad summary three point descriptive scale

- Working towards the expected level of attainment
- Achieving the expected level of attainment
- Working beyond the expected level of attainment

National Curriculum for Physical Education (2000)

- Four strands of learning: Acquiring and developing skill, Selecting and applying skills, Evaluating and improving performance and Knowledge and understanding of fitness and health
- Eight point scale plus exceptional performance

Developing teacher attitudes to assessment practice

Physical education teachers' awareness of the limitations of observation as a sole assessment strategy had increased and they were starting to recognise that alternatives to support this approach should be used. As I progressed through Stage 1 of the Doctoral programme, a gradual shift in attitude to assessment practice was becoming evident amongst the Physical Education teachers with whom I worked. In the early stages of the study, Physical Education teachers frequently argued that their "own professional judgement" was so well honed that they were very confident in their ability to reach a judgement about a pupils' progress based on ad hoc observation, often only undertaken once for each child!

However, it was noted that even during the lifetime of the study, from proposal (1998) to submission of stage 1 final report (2001), teachers were becoming more reluctant to state this position with confidence to both their trainees and the university based initial teacher training tutors. This was best evidenced in the context of data collection on assessment undertaken by trainees for their university assignments. They began to report that when questioned about approaches to assessment their mentors stated things like " I recognise we need to change our assessment practice" or " we are in the process of changing our assessment practice". Thus by the time of my deferral, it was becoming increasingly clear that the original focus of the study was shifting.

April 2003- September 2004

Deferral from Doctoral programme due to work commitments including leading Ofsted inspection, leading revalidation of initial teacher training courses and leading validation of new undergraduate course

Developments in assessment practice at a national level

Since deferring my EdD in April 2003, there has been significant progress in the area of assessment practice in schools. I perceive the reasons for this to be twofold. First, there has been significant activity in the publication of updated papers, which begin to address the complexity of the assessment process, in Physical Education in line with the National Curriculum for Physical Education (2000). This has opened the debate on the tension between validity and feasibility of assessment practice in school. However, this period of my own inactivity has also seen the development and adoption of the national Key Stage 3 strategy in state schools through out England and Wales. This, more than any other development, has very significant implications for my study; in particular work done on Assessment for Learning. At the time of deferring my study, whilst at the academic level work on Assessment for Learning was quite advanced (Assessment reform group,), at the practical implementation level in secondary schools it was in its infancy.

This climate change is due in no small part to the implementation of the Key Stage 3 Strategy, in which the approach to assessment, is underpinned by the work of the Assessment Reform Group (ARG), which includes such researchers such as, Paul Black, Richard Daugherty, Kathryn Ecclestone, John Gardner, Wynne Harlen, Mary James, Judy Sebba and Gordon Stobart, all of whom are leading researchers in this field. This implementation of research based practice into state education system has served to raise the profile of assessment practice in all subjects including Physical Education.

September 2004 - September 2005

Significantly refocused the study. The significant external changes, which occurred over the lifetime of this study, have led to a refocusing of the current research. Given the change in culture, as previously discussed, there was little or no value in examining the practice of a few individuals who are already identifying for themselves a need to change. Thus, rather than investigate the individuals who claim that they can make valid judgements based solely on teacher observation, a different method needed to be

undertaken. Whilst the evidence is often only anecdotal, in the new climate of assessment culture in Physical Education, the strong advocates for the sole use of teacher observation appear to have been silenced. Even if there is a feeling that over reliance on teacher observation is inherent in the departmental practice, individual teachers tend to articulate the view that they recognise this and are striving to change their practice.

Primary research question proposed in 2004

What approaches to assessment and reporting of pupil attainment and progress at Key Stage 3 are currently used in Physical Education and how far are these approaches satisfactory?

In particular, the research will examine the extent to which,

- Issues of reliability and validity in teacher assessment are considered.
- Teacher assessments of pupil attainment and progress are used to inform the end of Key Stage 3 summative reports to parents
- How Physical Education teachers make 'best-fit' judgements of pupil attainment at the end of Key Stage 3

Case study approach retained. More exploratory interpretative approach adopted, positivist hypothesis approach abandoned.
Focus to shift from teacher observation as the sole method for gathering evidence to an exploration of teacher assessment practice as used in Physical Education departments in one Physical Education initial teacher training partnership. Impact of changes made reported in the main thesis. Research questions further refined as research evolved.

June 2008 to September 2008 Deferral due to personal reasons
Unexpected illness and subsequent bereavement of father sadly delayed submission of this thesis for a further year.

Appendix Two

The dos and don't of assessment

Assessment Recording and Reporting in Physical Education Guidance for Teachers SCAA (1997)

DO

- Always distinguish between your ongoing evaluation, everyday formative assessment and your summative assessment of their attainment at the end of a period of time, e.g. a Key Stage.
- Focus upon the assessment criteria *planned for* in units of work and in your ongoing evaluation of their work.
- Spread the assessment among learning outcomes across the four aspects.
- Use the level descriptions to guide the *"pitch"* of the challenge in planned activities and this will help you make summative judgements later on.
- Work with pupils at target setting/pupil self-assessment strategies using the challenges incorporated in the QCA or own units of work and your lesson plans.
- Use all available information from the range of activities at different points in a key stage for the purpose of recording information on pupil progress and attainment.
- Make a rounded judgement and give a 'best-fit' level.
- Make effective use of assessment information for constructive feedback, future planning and reporting.
- Work and plan your way through each key stage, i.e. National Curriculum, schemes of work, units of work and lesson plans. It makes assessment easier.
- Use information from the level descriptions to write summative reports; they describe attainment
- Use all your notes and information at parent evenings.

DON'T

- Plan complicated recording sheets that take too long to fill in
- Plan to assess everything that moves
- Record more than you can use
- Level activities as though they were attainment targets, e.g. Level 4 Dance, Level 5 Games, etc.
- Add to your workload by writing about all sorts of things in summative reports.
- Get drawn into assessment of attitude, behaviour etc.

Appendix Three

Summary of key conclusions and implications from Harlen (2004a) review

Evidence in relation to the conditions that affect the reliability and validity of teachers' summative assessment led by Harlen (2004a)

Both high and medium weight evidence indicated the following:
There is bias in teachers' assessment (TA) relating to student characteristics, including behaviour (for young children), gender and special educational needs; overall academic achievement and verbal ability may influence judgement when assessing specific skills.

There is variation in the level of TA and in the difference between TA and standard tests or tasks that is related to the school. The evidence is conflicting as to whether this is increasing or decreasing over time. There are differences among schools and teachers in approaches to conducting TA.

There is no clear view of how reliability and validity of TA varies with the subject assessed. Differences between subjects in how TA compares with standard tasks or examinations results have been found, but there is no consistent pattern suggesting that assessment in one subject is more or less reliable than in another.

It is important for teachers to follow agreed procedures if TA is to be sufficiently dependable to serve summative purposes. To increase reliability, there is a tension between closer specification of the task and of the conditions under which it is carried out, and the closer specification of the criteria for judging performance.

The training required for teachers to improve the reliability of their assessment should involve teachers as far as possible in the process of identifying criteria so as to develop ownership of them and understanding of

the language used. Training should also focus on the sources of potential bias that have been revealed by research.

Teachers can predict with some accuracy their students' success on specific test items and on examinations (for 16 year-olds), given specimen questions. There is less accuracy in predicting 'A' level grades (for 18 year-olds).

Detailed criteria describing levels of progress in various aspects of achievement enable teachers to assess students reliably on the basis of regular classroom work.

Moderation through professional collaboration is of benefit to teaching and learning as well as to assessment. Reliable assessment needs protected time for teachers to meet and to take advantage of the support that others, including assessment advisers can give.

Conclusions

The implications of the findings of the review were explored through consultation with invited teachers, head teachers, researchers, representatives of teachers' organisations, of the Association for Achievement and Improvement through Assessment (AAIA), and of UK government agencies involved in national assessment programmes. Some points went beyond the review findings and are listed separately after those directly arising from the research evidence.

Implications for policy

When deciding the method, or combination of methods, of assessment for summative assessment, the shortcomings of external examinations and national tests need to be borne in mind.

The essential and important differences between TA and tests should be recognised by ceasing to judge TA in terms of how well it agrees with test scores.

There is a need for resources to be put into identifying detailed criteria that are linked to learning goals, not specially devised assessment tasks. This

will support teachers' understanding of the learning goals and may make it possible to equate the curriculum with assessment tasks.

It is important to provide professional development for teachers in undertaking assessment for different purposes that address the known shortcomings of TA.

The process of moderation should be seen as an important means of developing teachers' understanding of learning goals and related assessment criteria.

Implications for practice

Teachers should not judge the accuracy of their assessments by how far they correspond with test results, but by how far they reflect the learning goals.

There should be wider recognition that clarity about learning goals is needed for dependable assessment by teachers.

Teachers should be made aware of the sources of bias in their assessments, including the halo effect, and school assessment procedures should include steps that guard against such unfairness.

Schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers are sustained: for example, by protecting time for planning assessment, in-school moderation, etc.

Schools should develop an 'assessment culture' in which assessment is discussed constructively and positively, and not seen as a necessary chore (or evil).

Implications for research

There should be more studies of how teachers go about assessment for different purposes, what evidence they use, how they interpret it, etc.

The reasons for teachers' over-estimation of performance compared with moderators' judgements of the same performance, need to be investigated to find out, for instance, whether a wider range of evidence is used by the students' own teachers, or whether criteria are differently interpreted.

More needs to be known about how differences between schools influence the practice and dependability of individual teachers.

Since evaluating TA by correlation with test results is based on the false premise that they assess the same things, other ways need to be found for evaluating the dependability of TA.

There needs to be research into the effectiveness of different approaches to improving the dependability of TA, including moderation procedures.

Research should bring together knowledge of curriculum planners, learning psychologists, assessment specialists and practitioners to produce more detailed criteria that can guide TA

Additional points related to the review identified in consultation with users

It is important to consider the purpose of assessment in deciding the strengths and weaknesses of using teachers' assessment in a particular case. For instance, when assessment is fully under the control of the school and is used for informing pupils and parents of progress ('internal purposes'), the need to combine TA with other evidence (e.g. tests) may be less than when the assessment results are used for external purposes, such as accountability or the school or selection or certification of students.

There needs to be greater recognition of the difference between purposes of summative assessment and of how to match the way it is conducted with its purpose. For instance, the 'internal' assessment that is under the control of the school should not emulate the 'external' assessment, which has different purposes.

If tests are used, they should be reported separately from TA, which should be independent of the test scores.

There is evidence that a change in national assessment policy is due. The current system is not achieving its purpose. The recent report on comparability of national tests over time (Massey et al., 2003) concludes that TAs have shown less change in standards than the national tests. The authors state, 'National testing in its current form is expensive, primarily because of the external marking of the tests, and the time may soon come when it is thought that these resources may make a better contribution elsewhere' (Massey et al., 2003, p 239).

Improving teachers' formative assessment would also improve their summative assessment and so should be a part of a programme of

professional development aimed at enabling teachers' judgements to be used for summative purposes.

The role that pupils can take in their own summative assessment needs to be investigated and developed.

Any change towards greater use of TA in current systems where summative assessment is dominated by tests requires a major switch in resources from test development to supporting teacher-led assessment.

Change towards greater use of TA for summative purposes, requires a long-term strategy, with strong 'bottom-up' elements and provision for local transformations.

Appendix Four
Attainment target for NCPE (2000)

Level 1

Pupils copy, repeat and explore simple skills and actions with basic control and coordination. They start to link these skills and actions in ways that suit the activities. They describe and comment on their own and others' actions. They talk about how to exercise safely, and how their bodies feel during an activity.

Level 2

Pupils explore simple skills. They copy, remember, repeat, and explore simple actions with control and coordination. They vary skills, actions, and ideas and link these in ways that suit the activities. They begin to show some understanding of simple tactics and basic compositional ideas. They talk about differences between their own and others' performance and suggest improvements. They understand how to exercise safely, and describe how their bodies feel during different activities.

Level 3

Pupils select and use skills, actions and ideas appropriately, applying them with coordination and control. They show that they understand tactics and composition by starting to vary how they respond. They can see how their work is similar to and different from others' work, and use this understanding to improve their own performance. They give reasons why warming up before an activity is important, and why physical activity is good for their health.

Level 4

Pupils link skills, techniques and ideas and apply them accurately and appropriately. Their performance shows precision, control and fluency, and that they understand tactics and composition. They compare and comment on skills, techniques and ideas used in their own and others' work, and use this understanding to improve their performance. They explain and apply basic safety principles in preparing for exercise. They describe what effects exercise has on their bodies, and how it is valuable to their fitness and health.

Level 5

Pupils select and combine their skills, techniques and ideas and apply them accurately and appropriately, consistently showing precision, control and fluency. When performing, they draw on what they know about strategy, tactics and composition. They analyse and comment on skills and techniques and how these are applied in their own and others' work. They modify and refine skills and techniques to improve their performance. They explain how the body reacts during different types of exercise, and warm up and cool down in ways that suit the activity. They explain why regular, safe exercise is good for their fitness and health.

Level 6

Pupils select and combine skills, techniques and ideas. They apply them in ways that suit the activity, with consistent precision, control and fluency. When planning their own and others' work, and carrying out their own work, they draw on what they know about strategy, tactics and composition in response to changing circumstances, and what they know about their own and others' strengths and weaknesses. They analyse and comment on how skills, techniques and ideas have been used in their own and others' work, and on compositional and other aspects of performance, and suggest ways to improve. They explain how to prepare for, and recover from, the activities. They explain how different types of exercise contribute to their fitness and health and describe how they might get involved in other types of activities and exercise.

Level 7

Pupils select and combine advanced skills, techniques and ideas, adapting them accurately and appropriately to the demands of the activities. They consistently show precision, control, fluency and originality. Drawing on what they know of the principles of advanced tactics and compositional ideas, they apply these in their own and others' work. They modify them in response to changing circumstances and other performers. They analyse and comment on their own and others' work as individuals and team members, showing that they understand how skills, tactics or composition and fitness relate to the quality of the performance. They plan ways to

improve their own and others' performance. They explain the principles of practice and training, and apply them effectively. They explain the benefits of regular, planned activity on health and fitness and plan their own appropriate exercise and activity programme.

Level 8

Pupils consistently distinguish and apply advanced skills, techniques and ideas, consistently showing high standards of precision, control, fluency and originality. Drawing on what they know of the principles of advanced tactics or composition, they apply these principles with proficiency and flair in their own and others' work. They adapt it appropriately in response to changing circumstances and other performers. They evaluate their own and others' work, showing that they understand the impact of skills, strategy and tactics or composition, and fitness on the quality and effectiveness of performance. They plan ways in which their own and others' performance could be improved. They create action plans and ways of monitoring improvement. They use their knowledge of health and fitness to plan and evaluate their own and others' exercise and activity programme.

Exceptional Performance

Pupils consistently use advanced skills, techniques and ideas with precision and fluency. Drawing on what they know of the principles of advanced strategies and tactics or composition, they consistently apply these principles with originality, proficiency and flair in their own and others' work. They evaluate their own and others' work, showing that they understand how skills, strategy and tactics or composition, and fitness relate to and affect the quality and originality of performance. They reach judgements independently about how their own and others' performance could be improved, prioritising aspects for further development. They consistently apply appropriate knowledge and understanding of health and fitness in all aspects of their work.

Appendix Five

Programme of study: PE Key Stage 3

Knowledge, Skills and Understanding

Teaching should ensure that, when **evaluating and improving performance**, connections are made between **developing, selecting and applying skills, tactics and compositional ideas**, and **fitness and health.**

Acquiring and developing skills

1 Pupils should be taught to:

a refine and adapt existing skills

b develop them into specific techniques that suit different activities and perform these with consistent control.

Selecting and applying skills, tactics and compositional ideas

2 Pupils should be taught to:

a use principles to plan and implement strategies, compositional and organisational ideas in individual, pair, group and team activities

b modify and develop their plans

c apply rules and conventions for different activities.

Evaluating and improving performance

3 Pupils should be taught to:

a be clear about what they want to achieve in their own work, and what they have actually achieved

b take the initiative to analyse their own and others' work, using this information to improve its quality.

Knowledge and understanding of fitness and health

4 Pupils should be taught:

a how to prepare for and recover from specific activities

b how different types of activity affect specific aspects of their fitness

c the benefits of regular exercise and good hygiene

d how to go about getting involved in activities that are good for their personal and social health and well being.

Appendix Six
Questionnaire Schedule (Method A)

Sheffield Hallam University
School of Education

Name: School:

LEA (if applicable): M\F

1. Please indicate your level of usage of each of the following evidence collection tools, in your assessment strategy for Key Stage 3 Physical Education, by circling a number on each scale.

5= frequently used 1= Never used

A Teacher observation 5 4 3 2 1
Please outline your reasons.

B Written tests 5 4 3 2 1
Please outline your reasons.

C Peer assessment 5 4 3 2 1
Please outline your reasons.

D Task cards 5 4 3 2 1
Please outline your reasons.

E Video recording 5 4 3 2 1
Please outline your reasons.

2. Do you have a preferred choice of evidence collection tool?

Yes No (please circle)

If yes, please indicate which, and give reasons for your choice.

3. Please indicate your level of confidence that *your* Year 9 summative reports are accurate in terms of reporting pupil achievement and progress against *all four* strands contained in the attainment target for Physical Education at the end of Key Stage 3.

Please circle one number on the scale below.

5= High level of confidence. 1= Low level of confidence.

5 4 3 2 1

Thank you for taking the time to complete this questionnaire.

Please return the completed questionnaire to Diane Burkinshaw Sheffield Hallam University, School of Sport and Leisure management, Collegiate Hall, Collegiate Crescent Sheffield S10 2BP

Appendix Seven
Tasks for school placement revised (2001)

To complete tasks 1 and 2 you are required to conduct a semi-structured (guided) interview with your school-based mentor. You must produce a written account of the main findings of your interview, which both you and your mentor must sign to confirm as an accurate record. A summary of your lesson observation notes and the interview account must be submitted as an appendix to your assignment for this unit, (PYSPPE3-1).

1. Discuss with your school mentor the types of assessment used in the Physical Education department to gather evidence of pupil attainment and progress at Key Stage 3
2. Discuss with your mentor how the issues of objectivity, validity, and reliability in assessment are addressed within the Physical Education department.
3. When you have completed your interview with your mentor, you are required to observe TWO lessons, to see the extent to which your mentor implements issues from your discussion into their practice.

You should use Task 11.4 p.167, in Capel, S (1997) " Learning to Teach Physical Education in the Secondary school to structure your observations, which is detailed below.

During the observations and extra curricular activity, draw up a list of the following point from your observations

- examples of **methods** used for assessing pupils. This answers the question of how pupils are assessed (for example, observation of performance, listening to answers to questions, writing down scores/comments, written comments by pupils or assignments.
- examples of **what** the teacher is assessing I attitudes, planning, performance evaluation cooperation
- examples of **who** is doing the assessment. Is it always the teacher?

- examples of **why** the assessment is being applied. Is it to give feedback to the pupils/parents/governors/others. Is it to motivate? Is it to identify the best performers? Any other reasons?
- examples of how pupils are given the results of assessment. Is it through an informal process such as a brief comment giving positive or negative feedback? Is it through a mark given for a specific performance or evaluation? Any other ways?

Source: Capel (1997) p.167).

Task One and Two

Interview questions

The following questions MUST be asked. However, you may also use supplementary questions as required.

1. What types of assessment are used in the PE department to gather evidence of pupil attainment and progress?
2. How do you address the issue of objectivity in assessment within the PE department?
3. How do you address the issue of validity in assessment within the PE department?
4. How do you address the issue of reliability in assessment within the PE department?
5. How do you record pupils' progress?
6. How do you report pupils' progress to parents?

Appendix Eight
Interview schedule

Departmental issues

1. Approaches to KS3?
2. Types of teacher assessment?
3. Sharing learning objectives?
4. Planned and systematic approach?
5. Internal moderation systems?
6. Approaches to validity and reliability?
7. How levels are decided upon?
8. Formative and summative approaches?
9. Departmental school policy on assessment at KS3?

Whole school issues

1. Whole school policy on assessment at KS3?
2. Staff development time available?
3. Staff development courses available?

Appendix Nine
Ofsted (2003b) Good Assessment Practice in Physical Education

Features of effective assessment

1. Effective assessment in Physical Education is integral to teaching and learning.
2. The clarity of teachers' planning is also central to good assessment. Short- and medium-term planning, setting out with precision what it is that teachers want pupils to know, understand and do, ensures strong and essential links between planning and assessment.
3. Clear rationale for the subject that defines what is to be learned about movement and its application.
4. Teachers share these intentions with pupils to enhance understanding of what is to be learned at different stages throughout a lesson and the unit of work.

On-going assessment

1. A well-developed policy that is explicit about assessment purposes and procedures and provides good guidance.
2. Teachers ensure precise, shared, learning objectives are used to check pupils' progress at different stages throughout lessons.
3. Careful observation of pupils' responses to tasks, identifying strengths, errors and misconceptions. Use this information to intervene and provide specific feedback to guide pupils towards improvement.
4. Use of demonstrations and different types of well-focused questioning of pupils' knowledge and understanding is reinforced through pupils' practical responses
5. Teachers concentrate on the needs of individual pupils rather than simply completing the lesson.
6. Thorough account is taken of the quality of individual pupils' responses to the tasks set, work is differentiated to cater for the more able or the less able pupils and all pupils are set tasks appropriate to their previous performance.

7. The use of targets is becoming a regular feature in Physical Education at best these are specific, realistic and achievable.
8. Using assessment to improve provision e.g. analyse results to look for issues that can then be resolved e.g. fragmented curriculum
9. The exchange of assessment information between primary and secondary schools remains a challenge for all schools. To effectively meet this, some secondary departments are beginning to construct a 'baseline level' for new Year 7 pupils using National Curriculum levels. This is intended to show the progress pupils will have made by the end of the key stage.
10. Increasingly, schools are using data to compare the achievements of boys and girls and are using data to provide an action plan for raising achievement if either group lags behind.

Involving pupils in the assessment process

1. Providing clearly structured opportunities to ensure that pupils are involved in the assessment process and take some responsibility for assessing their own performance against known and understood criteria. This self-assessment takes different forms.
2. On a day-to-day basis, teachers set tasks or ask questions that engage pupils in direct observation or analysis of their own and each other's physical performances, and create opportunities for them to discuss and evaluate these performances helping them to identify areas for improvement.
3. The most effective departments ensure that pupils have well-structured opportunities to develop their observation and evaluation skills across a key stage. At the end of each term, pupils at one school complete a personal performance diary recording their perceptions of their progress and achievement in particular aspects of the PE curriculum.

Standardisation and moderation

1. Unit planning and assessment is linked to the National Curriculum programme of study;

2. Precise learning objectives are described in language that pupils understand.
3. Teachers have an agreed view on what constitutes performance at each level across all aspects of the programme of study and areas of activity.
4. They achieve standardisation by discussing pupils' work to establish criteria for performance at each level.
5. These end-of-unit assessments are used cumulatively to determine attainment against National Curriculum levels
6. Levels are recorded using +/- to indicate subtle differences between pupils.
7. Internal moderation procedures are used to help standardise judgements and expectations in order to moderate the assessment of non-examination work across both key stages.

Appendix Ten

Framework for analysis

Analysis against Ofsted and Harlen framework Year

School Names					
Assessment Purposes					
Summative					
Formative					
Assessment Types					
Formal					
Informal					
Assessment methods					
Teacher observation					
Peer assessment					
Written assessment					
Self reflection					
Target setting					
Target setting against levels ongoing					
Formal Levelling					
Validity considered					
Reliability considered					
Reliance on teachers' Professional Judgement justified					

Ofsted **Effective assessment in PE seen where evidence of**	**EPPI** **Dependability increased where there is evidence of**					
Conditions the affect dependability of assessment						
Well developed assessment policy, explicit guidance about the purposes and procedures for assessment						
	Awareness of potential teacher bias, due to irrelevant factors behaviour, gender, SEN					
	Whole school action on assessment, eg PPA time					
	Whole school positive culture for assessment, eg shared discussions					
Ongoing assessment						
Assessment is integral to teaching and learning not bolt on	Accuracy of assessment linked to learning goals not test result					
Clarity and precision in planning for assessment, short medium term, what teacher wants pupils to know do and understand at each stage	Clearly defined assessment tasks linked to learning goals					
Precise shared learning objectives used to check pupil progress	Clearly articulated learning goals					
	Clearly defined assessment tasks linked to learning goals					
Careful observation of pupil responses to task used to						

provide specific feedback to guide pupils towards improvement						
Use of feedback and target setting to facilitate progress						
Involving pupils in the assessment process						
	Detailed assessment criteria linked to learning goals					
Opportunities for Pupil peer and self assessment against known and understood criteria						
Opportunities to observe and evaluate each others work to identify areas for improvement						
Shared criteria for assessment in language pupils understand	Pupils understand assessment criteria and know what they have to do to meet them					
Progress and attainment recorded in pupil progress file						
Standardisation and moderation						
	Progressive levels of attainment defined					
Shared teacher understanding of NC levels of attainment	Opportunities provided for teachers to share 'good practice' in assessment					
End of unit assessment used cumulatively to determine achievement against NC levels of attainment						
Planning and assessment linked to NC programme of study	Detailed but generic assessment criteria which allow evidence collected					

	from a range of class work					
	Progressive levels of attainment defined					
Standardisation through discussion of pupil work to establish criteria for performance at every level						
Levels recorded + / - to show subtle differences between pupils						

X = No evidence

☆ = Some evidence

☆☆ = Significant evidence

☆☆☆ = Part of teachers everyday practice

Appendix Eleven

Key criteria used in methodology in relation to summative assessment in Physical Education from Harlen (2004a)

Implications for practice

Implication from EPPI review	Interpretation for present research
Teachers should not judge the accuracy of their assessments by how far they correspond with test results, but by how far they reflect the learning goals.	Accuracy of assessment judged by extent to which reflect learning goals
There should be wider recognition that clarity about learning goals is needed for dependable assessment by teachers.	Clarity in learning goals increases dependability of assessment
Schools should take action to ensure that the benefits of improving the dependability of the assessment by teachers is sustained: for example, by protecting time for planning assessment, in-school moderation,.	whole school commitment to providing time for in school moderation, planning
Schools should develop an 'assessment culture' in which assessment is discussed constructively and positively, and not seen as a necessary chore (or evil).	Assessment culture discussion of assessment in positive climate

Appendix Twelve

Summary of methods: Harlen (2004a)

A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes.

The review methodology followed the procedures devised by the Evidence for Policy and Practice Information and Co-coordinating Centre (EPPI-Centre), and the Review Group received the technical support of the EPPI-Centre. Criteria were defined for guiding a wide-ranging search for studies that dealt with some form of summative assessment conducted by teachers, involving students in school in the age range 4 to 18, and reporting on the validity and/or reliability of methods used. Bibliographic databases and registers of educational research were searched online as were relevant online journals, with other journals and back numbers of those only recently put online being searched by hand. Other studies were found by scanning the references lists of already-identified reports, making requests to members of relevant associations and other review groups, and using personal contacts.

All studies identified in these ways were screened, using inclusion and exclusion criteria, and the included studies were then key worded, using the *Core Key wording Strategy* (EPPI-Centre, 2002a) and additional keywords specific to the context of the review. Keywords were used to produce a map of selected studies. Detailed data extraction was carried out online independently by two reviewers who then worked together to reach a consensus, using the EPPI-Reviewer (*Review Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research* (EPPI-Centre, 2002b)). Review-specific questions relating to the weight of evidence of each study in the context of the review were used in addition to those of the EPPI-Reviewer. Judgements were made as to the weight of evidence relevant to the review provided by each study in relation to methodological soundness, appropriateness of the study type and relevance of the focus to the review questions.

The structure for the synthesis of evidence from the in-depth review was based on the extent to which the studies were concerned with reliability or validity of the assessment. Despite the difficulty in making a clear distinction between these concepts, and their inevitable interdependence, it was possible to designate each one as providing evidence *primarily* in relation to reliability or *primarily* in relation to validity. Evidence in relation to the conditions affecting reliability or validity was drawn together separately. In the synthesis and discussion, reference was made to the weight of evidence provided by each study.

Potential users of the review were involved in several ways: providing advice as members of the review group; providing information about studies through personal contact; participating in keywording and in data extraction; and through a consultation seminar on implications of the draft findings of the review attended by a number of policy and practitioner users.

Identification of studies

The search resulted in a total of 431 papers being found. Of these, 369 were excluded, using exclusion criteria. Full texts were obtained for 48 of the remaining 62 papers, from which a further 15 were excluded, and two sets of papers (three in one case and two in the other) were linked as they reported on the same study. This left 30 studies after key wording. All of these were included in the in-depth review.

Systematic map

The 30 studies included in the in-depth review were mapped in terms of the EPPI-Centre and review-specific keywords. All were written in the English language: 15 were conducted in England, 12 in the United States and one each in Australia, Greece and Israel. All studies were concerned with students between the ages of 4 and 18. Of the 30, 11 involved primary school or nursery students (aged 10 or below) only, 13 involved secondary students (aged 11 or above) only, and six were concerned with both primary and secondary students. There was no variation across educational settings in terms of whether the study focus was on reliability or validity, but there were slightly more evaluations of naturally-occurring situations in primary

schools. Almost all studies set in primary and nursery schools involved assessment of mathematics and a high proportion related to reading. At the secondary level, studies of assessment of mathematics and 'other' subjects (variously concerned with foreign languages, history, geography, Latin and bible studies) predominated.

Eighteen studies were classified as involving assessment of work as part of, or embedded in, regular activities. Three were classified as portfolios, two as projects and nine were either set externally or set by the teacher to external criteria. The vast majority were assessed by teachers, using external criteria. The most common purpose of the assessment in the studies was for national or state-wide assessment programmes, with six studies related to certification and another six to informing parents (in combination with other purposes). As might be expected in the context of summative assessment, most research related to the use of external criteria by teachers, with little research on student self-assessment or teachers using their own criteria.

Appendix Thirteen
Examples of Raw data

Campion Comprehensive YEAR: 2000

Q - Discuss with your mentor / PE teacher the types of assessment used in the PE department to gather evidence of pupil attainment and progress.

A – The types of assessment used in the PE department for KS3 are practical based. Pupils are observed by the PE teacher and give a level for each activity they do. These levels are taken straight from the PE National Curriculum document and are recorded in an assessment sheet. There are four different levels, which can be issued for one activity, these are acquiring and developing skills, selecting and applying skills, tactics and compositional ideas, evaluating and improving performance and knowledge and understanding of fitness. Within years seven, eight and nine pupils will have an interim and a final grade.

Pupils are also given a grade for behaviour and effort, which ranges from A being excellent to D, which is need for improvement. The final grades for the level descriptors and the grades given for effort and behaviour are recorded is a teacher's bromcom. These grades are sent to the school office and are printed off for pupil's reports.

In addition to this the PE department is introducing a multiple choice question sheet for the different activities at KS3. Pupils will be given a grade for each activity and this will be recorded, to highlight if learning has occurred.

During KS4 the pupils have the opportunity to take part in the Junior Sports Leader Award (JSLA). This scheme requires the pupils to complete a training programme for each unit, a home project and a written plan for an activity session. The fitness programme is marked and given back to the pupil along with a feedback sheet, this sheet is photocopied and the PE

department keeps one. This work will determine whether a pupil will achieve a fail, pass, merit or distinction.

If a pupil takes GCSE PE as part of their options they will be assessed in their course work, written training programmes, tests, exams and practically with the use of videos etc. The GCSE syllabus is divided into four sections, it is required that pupils have a knowledge of the rules and regulations of particular activities. Campion have designed a question sheet, which tests a pupil's knowledge. Section D is divided into two parts; one and two both are given a mark out of ten. Section D1 concerns analyzing performance and section D2 concerns improving performance.

To gain an overall grade for GCSE PE coursework, practical assessment and a written exam will determine the graded achieved. For the PE teacher to determine an estimated grade, mock tests, previous coursework and practical observations are used.

Q - Discuss with your mentor / PE teacher how the issues of objectivity, validity and reliability in assessment are addressed in the PE department in your school.

A – At Campion Comprehensive School it is recognized that all the PE teachers in the department have to be working towards the same objectives and outcomes. Assessment within the department should be progressive and constantly evaluated, to identify if the right assessment has occurred and to check that the right methods are achieving these views.

The National curriculum requests that pupils are to be assessed at the end of every key stage, but to track grades it is necessary for Campion to record results throughout the school year, for every year group. This helps to evaluate learning and report back to parents.

Within the whole school there is a scheme called 'Performance Management'. This assesses the objectivity, validity and reliability of a teacher's performance. Within a faculty one member of staff, normally the head of department has to evaluate a teacher's performance. As part of this scheme individual teachers have to set themselves targets for the year and aim to achieve these.

A step up from this, all schools are involved in Ofsted inspections. Teachers have to be observed in up to three of their lessons. As a department they are also evaluated concerning their assessment formats, record keeping and the general running of the department.

Q - Discuss with your mentor / PE teacher the methods used to record any assessment information collected. Obtain a copy of any recording sheets or proformas used in your school.

A – At KS3 pupils are given a level describing four areas of the curriculum. These involve acquiring and developing skill, selecting and applying skills, tactics and compositional ideas, evaluating and improving performance and knowledge and understanding of fitness. A level is given for every activity involving the four areas. The department, to record these levels, has produced a format. Each pupil will have a sheet, which provides information up until year nine. The department finds their format the most manageable and organized way of recording assessment. The record sheets are filed into PE groups and can be used by all members of staff.

At GCSE level the department uses the AQA format to record levels and results. The department has added to this format by identifying a section titled fitness programme, which allocates marks for planning performing and evaluating. Another section highlights exam results, percentage givens for section A and B, an estimated grade and an overall grade.

Course work and task marks, are recorded by individual teachers in their planner for their own overview of learning.

Q – Discuss with your mentor / PE teacher the system used in the school to report to parents. Obtain a copy of any documentation used.

A – Annual reports are a statutory requirement, PE teachers are required to record judgements against the level descriptions in PE at the end of KS3. At Campion Catholic High School all of the teachers have to type up reports for all of their pupils at the end of each year. Examples have been given from years seven, eight and nine, which are based on the activities they are involved in, attitude and behaviour, organization, class work and effort, there is also a section which allows the teacher to give an overall comment about the pupil.

In years ten and eleven GCSE PE the teacher is required to word-process a blurb about individual pupils, which is sent to the office to be typed into their National Records of Achievement (NRA).

The school has obviously developed their own strategies to report to parents. As a guide the PE department refer to 'Physical education Assessment, Recording and Reporting at Key Stages 1 to 4', produced by the Physical Education Association.

Q – Complete the task marked * in your handout defined by Capel (1997) in her chapter on 'Assessment in Learning to Teach in PE in Secondary School', Routledge.

A – Lesson 1 – GCSE PE

This lesson was based on the pupils representing a short presentation about the eleven components of fitness. Each presentation involved a different component of fitness. While all of the class watched the presentations, the

teacher observed and noted down some positive and negative qualities. After every group presented, feedback was generated involving the whole class, this involved the pupils in the assessment process. The teacher asked the class and the groups presenting questions concerning the component and activity and how this could relate and help when planning a training programme (this is relevant for a piece of course work in year eleven).

The pupils were required to support the presentation with a word-processed handout, which was collected by the teacher and used to support her observations. The teacher was assessing how well the pupils work in groups, their ability to research the topic, their planning skills and their presentation skills.

The reasons for the teacher assessing this unit were predominately to assess whether learning had occurred and to reinforce what had been taught in previous weeks. Involving the pupils in these presentations was also a different way to involve the pupils in the lesson, which in this case seemed to motivate the pupils to learn.

The pupils were given a grade and a merit certificate to support their effort.

Lesson 2 – Yr7 Gymnastics

This lesson was based on flight on and off apparatus. The teacher, assessing the ability of individual pupils, frequently observed the lesson. This was necessary because some pupils could progress on to more difficult tasks and others were more comfortable with staying on one piece of apparatus doing simple movements.

The teacher asked the class questions concerning the dos and don't of gymnastics – what happens if you look at your toes? Answer – you fall on your nose.

The reason for the ongoing assessment was due to safety aspects of gymnastics and also to help pupil's progress and to build confidence to learn

new moves. The teacher never forced all of the pupils to perform the same movement; this would only scare the pupils and produce negative thoughts about gymnastics.

The pupils were praised constantly during the lesson and given feedback to support their development.

Raw Data

Croft School YEAR 2006

TASK 1

Q – What types of assessment used in the PE department to gather evidence of pupil attainment and progress.

A – The main strategy for assessment within the school is the observation of performance by teachers. A secondary assessment strategy used is question and answer.

Q and A is used by all teachers, even if it wasn't planned for. Q and A occurs all the time during every lesson the check for example: to check for understanding of the task, to ask what was good about a performance, to ask for ideas as to how a task can be completed, to ask for an understanding of a warm up…the list could go on.

Peer observation and feedback is used, especially within dance. This links to the evaluating and improving aspect of the national curriculum. So, with peer evaluation pupils look to see what is good about a performance and what could be improved.

Formative assessment occurs all the time. Assessment is on-going during every lesson, so during lesson the teacher can say who is performing the best and who is struggling. Teachers are constantly thinking about what levels pupils are at, and in most lessons they will have a rough idea as to what level pupils are at. In addition to this, the teachers do a formative assessment of all the pupils half way through a block of work, as well as a formative assessment in the final week(s) of the block of work.

Summative reviews occur at the end of a module of work. These reviews are done by giving the pupils a end of key stage descriptor (EKSD), which

is a level of one (being poor) to 8 (being sporting excellence). An average of all of the EKSD received by the pupils are recorded on the school database and used in the pupils report.

Self assessment occurs within the school. The use of colour coded assessment strands worded so that they are easy for the pupils to understand enables the pupils to look at the assessment criteria and decide what level they believe they are at and what level they believe they can reach. The help with this a year 7 PE and Games booklet has been produced in which the pupils record what they have learned and what level they believe they are at.

TASK 2-

Discuss with your mentor / PE teacher how the issues of objectivity, validity and reliability in assessment are addressed in the PE department in your school.

Q - How do you address the issue of objectivity in assessment within the PE department?

A - Objectivity – This occurs through the use of EKSD levels. These have been re-written in line with the revisited assessment policy at Key Stage 3. For the ESKD there are set criteria, which all the teaching staff in PE have copies of. In addition to this the levels are sport specific. Therefore assessment is objective for each sport as there is set criteria for each sport to follow. This means that teachers are assessing in relation to set criteria, rather than interpreting performance.

Q - How do you address the issue of validity in assessment within the PE department?

A - Validity – Teachers must assess what they say they are going to assess. Therefore the teachers assess their learning outcomes; they check that the

children have done what they are meant to do. The line manager assesses the class teachers' assessment; this keeps a check on the class teachers assessment skills.

Q - How do you address the issue of reliability in assessment within the PE department?
A - Reliability – The ESKD levels were re-written and every member of staff were given a copy, so they now all have the same assessment criteria to assess the pupils with. Therefore the pupils should get the correct level, they should get the same level regardless to which teacher assesses them.

At key stage four there is a moderation day where the moderator comes into the school to check that the teachers are giving the correct levels. A mock moderation is held by the school where all the teachers will come together with the pupils to assess each other's pupils to see if they agree with the levels given out.

Q - How do you record pupils' progress?
A – Teachers record their grades on the school database. In year 7, pupils record their progress in their PE booklets. Going to be spread to other years

Q – How do you report pupils' progress to parents? Obtain a copy of any documentation used.

A – Pupils are given a report to take home to their parents once every school year (usually at the end of the school year). In addition to this within all years 7 – 11 they have a parents evening once a year, where the parents have the opportunity to see each subject teacher. In years 7 and 8 there is a parents/tutor meeting, thus being a meeting between the pupils' form/registration tutor and the pupils parent(s). Finally in years 12 and 13 they receive a progress review report twice a year. Year 12 and 13 also have the parents evening once a year, with year 13 pupils having a second parents evening where higher education is discussed.

TASK 3

Q – Observe two lessons
Structure the observations around the task marked * in your handout defined by Capel (1997) in her chapter on 'Assessment" in Learning to Teach in PE in Secondary School', Routledge.

A –
Lesson 1

Methods – Q and A was used the most during the lesson to assess the pupils, the other form of assessment was teacher observation of performance, and attitudes of the pupils within the lesson.

What – the teacher was assessing the following:
Pupil performance – How well they were performing
Attitude – What the pupils' attitude was towards the task , other pupils and towards the teacher
On task – The teacher was consistently assessing whether the pupils were on task i.e. were all the pupils doing what they were asked to do.
Understanding – In addition to assessing whether the pupils were on task, the teacher also assessed whether the pupils understood the task, and whether they understood why they were doing it i.e. did they understood that creating width helped attacking play?

Who – The teacher did most of the assessing. However, at one stage the pupils gathered around one square to watch a group. The pupils watching were then asked to say what was good about the performance and how they could improve. Therefore the pupils were also used to assess the performance.

Why – They assessed the group for the following reasons:

Motivate – It was an extremely cold day, which de-motivates pupils. By giving constant feedback on performance (a simple well done, that was good.) this can help keep children motivated.
Check for understand – The pupils was assessing to see that the pupils knew what they were doing, if they understood the task.
Evaluate and improve – The peer assessment enables the pupils to think about how they could improve the performance.

How – Positive reinforcement

Lesson 2

Methods - Generally teacher observation and Q and A. Peer assessment was used when pupils watched a group demonstrate.

What – The teacher was assessing the following:
Pupil performance – How well they were performing
Attitude - - What the pupils' attitude towards the task, other pupils and towards the teacher.
On task – The teacher was consistently assessing whether the pupils were on task i.e. were all the pupils doing what they were asked to do.
Understanding – In addition to assessing whether the pupils were on task, the teacher also assessed whether the pupils understood the task, and whether they understood why they were doing it i.e. did they understood that creating width helped attacking play?

Who – Again, generally the teacher did most of the assessing; however, peer assessment was done when peers watched the group demonstrating.

Why – The group were assessed for the following reasons: Motivation, Check understanding and to improve performances as well as knowledge and understanding of the tasks.

How – Positive reinforcement, Q and A

I confirm that this is an accurate account of the discussion with my trainee.

Name of Mentor:
XXXXXXX

Signed XXXXX Date 2006